Preface

Welcome to *College Physics*, an OpenStax resource. This textbook was written to increase student access to high-quality learning materials, maintaining highest standards of academic rigor at little to no cost.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 20 books for college and AP courses used by hundreds of thousands of students. Our adaptive learning technology, designed to improve learning outcomes through personalized educational paths, is being piloted in college courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax Resources

Customization

College Physics is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit your book page on openstax.org for more information.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. Since our books are web based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on openstax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past errata changes on your book page on openstax.org.

Format

You can access this textbook for free in web view or PDF through openstax.org, and in low-cost print and iBooks editions.

About *College Physics*

College Physics meets standard scope and sequence requirements for a two-semester introductory algebra-based physics course. The text is grounded in real-world examples to help students grasp fundamental physics concepts. It requires knowledge of algebra and some trigonometry, but not calculus. *College Physics* includes learning objectives, concept questions, links to labs and simulations, and ample practice opportunities for traditional physics application problems.

Coverage and Scope

College Physics is organized such that topics are introduced conceptually with a steady progression to precise definitions and analytical applications. The analytical aspect (problem solving) is tied back to the conceptual before moving on to another topic. Each introductory chapter, for example, opens with an engaging photograph relevant to the subject of the chapter and interesting applications that are easy for most students to visualize.

Concepts and Calculations

The ability to calculate does not guarantee conceptual understanding. In order to unify conceptual, analytical, and calculation skills within the learning process, we have integrated Strategies and Discussions throughout the text.

Modern Perspective

The chapters on modern physics are more complete than many other texts on the market, with an entire chapter devoted to medical applications of nuclear physics and another to particle physics. The final chapter of the text, “Frontiers of Physics,” is devoted to the most exciting endeavors in physics. It ends with a module titled “Some Questions We Know to Ask.”

Key Features

Modularity

This textbook is organized as a collection of modules that can be rearranged and modified to suit the needs of a particular professor or class. That being said, modules often contain references to content in other modules, as most topics in physics cannot be discussed in isolation.

Learning Objectives

Every module begins with a set of learning objectives. These objectives are designed to guide the instructor in deciding what content to include or assign, and to guide the student with respect to what he or she can expect to learn. After completing the module and end-of-module exercises, students should be able to demonstrate mastery of the learning objectives.

Call-Outs

Key definitions, concepts, and equations are called out with a special design treatment. Call-outs are designed to catch readers' attention, to make it clear that a specific term, concept, or equation is particularly important, and to provide easy reference for a student reviewing content.

Key Terms

Key terms are in bold and are followed by a definition in context. Definitions of key terms are also listed in the Glossary, which appears at the end of the module.

Worked Examples

Worked examples have four distinct parts to promote both analytical and conceptual skills. Worked examples are introduced in words, always using some application that should be of interest. This is followed by a Strategy section that emphasizes the concepts involved and how solving the problem

relates to those concepts. This is followed by the mathematical Solution and Discussion.

Many worked examples contain multiple-part problems to help the students learn how to approach normal situations, in which problems tend to have multiple parts. Finally, worked examples employ the techniques of the problem-solving strategies so that students can see how those strategies succeed in practice as well as in theory.

Problem-Solving Strategies

Problem-solving strategies are first presented in a special section and subsequently appear at crucial points in the text where students can benefit most from them. Problem-solving strategies have a logical structure that is reinforced in the worked examples and supported in certain places by line drawings that illustrate various steps.

Misconception Alerts

Students come to physics with preconceptions from everyday experiences and from previous courses. Some of these preconceptions are misconceptions, and many are very common among students and the general public. Some are inadvertently picked up through misunderstandings of lectures and texts. The Misconception Alerts feature is designed to point these out and correct them explicitly.

Take-Home Investigations

Take Home Investigations provide the opportunity for students to apply or explore what they have learned with a hands-on activity.

Things Great and Small

In these special topic essays, macroscopic phenomena (such as air pressure) are explained with submicroscopic phenomena (such as atoms bouncing off walls). These essays support the modern perspective by describing aspects of modern physics before they are formally treated in later chapters. Connections are also made between apparently disparate phenomena.

Simulations

Where applicable, students are directed to the interactive PHeT physics simulations developed by the University of Colorado. There they can further explore the physics concepts they have learned about in the module.

Summary

Module summaries are thorough and functional and present all important definitions and equations. Students are able to find the definitions of all terms and symbols as well as their physical relationships. The structure of the summary makes plain the fundamental principles of the module or collection and serves as a useful study guide.

Glossary

At the end of every module or chapter is a Glossary containing definitions of all of the key terms in the module or chapter.

End-of-Module Problems

At the end of every chapter is a set of Conceptual Questions and/or skills-based Problems & Exercises. Conceptual Questions challenge students' ability to explain what they have learned conceptually, independent of the mathematical details. Problems & Exercises challenge students to apply both concepts and skills to solve mathematical physics problems. Online,

every other problem includes an answer that students can reveal immediately by clicking on a “Show Solution” button.

In addition to traditional skills-based problems, there are three special types of end-of-module problems: Integrated Concept Problems, Unreasonable Results Problems, and Construct Your Own Problems. All of these problems are indicated with a subtitle preceding the problem.

Integrated Concept Problems

In Integrated Concept Problems, students are asked to apply what they have learned about two or more concepts to arrive at a solution to a problem. These problems require a higher level of thinking because, before solving a problem, students have to recognize the combination of strategies required to solve it.

Unreasonable Results

In Unreasonable Results Problems, students are challenged to not only apply concepts and skills to solve a problem, but also to analyze the answer with respect to how likely or realistic it really is. These problems contain a premise that produces an unreasonable answer and are designed to further emphasize that properly applied physics must describe nature accurately and is not simply the process of solving equations.

Construct Your Own Problem

These problems require students to construct the details of a problem, justify their starting assumptions, show specific steps in the problem’s solution, and finally discuss the meaning of the result. These types of problems relate well to both conceptual and analytical aspects of physics, emphasizing that physics must describe nature. Often they involve an integration of topics from more than one chapter. Unlike other problems, solutions are not provided since there is no single correct answer.

Instructors should feel free to direct students regarding the level and scope of their considerations. Whether the problem is solved and described correctly will depend on initial assumptions.

Additional Resources

Student and Instructor Resources

We've compiled additional resources for both students and instructors, including Getting Started Guides, an instructor solution manual, and PowerPoint slides. Instructor resources require a verified instructor account, which can be requested on your openstax.org log-in. Take advantage of these resources to supplement your OpenStax book.

Partner Resources

OpenStax Partners are our allies in the mission to make high-quality learning materials affordable and accessible to students and instructors everywhere. Their tools integrate seamlessly with our OpenStax titles at a low cost. To access the partner resources for your text, visit your book page on openstax.org.

About the Authors

Senior Contributing Authors

Paul Peter Urone, Professor Emeritus at California State University, Sacramento
Roger Hinrichs, State University of New York, College at Oswego

Contributing Authors

Kim Dirks, University of Auckland
Manjula Sharma, University of Sydney

Reviewers

Matthew Adams, Crafton Hills College, San Bernardino Community College District
Erik Christensen, South Florida Community College
Douglas Ingram, Texas Christian University
Eric Kincanon, Gonzaga University
Lee H. LaRue, Paris Junior College
Chuck Pearson, Virginia Intermont College
Marc Sher, College of William and Mary
Ulrich Zurcher, Cleveland State University

Introduction to Oscillatory Motion and Waves

class="introduction"

There are at least four types of waves in this picture—only the water waves are evident. There are also sound waves, light waves, and waves on the guitar strings. (credit: John Norton)

What do an ocean buoy, a child in a swing, the cone inside a speaker, a guitar, atoms in a crystal, the motion of chest cavities, and the beating of hearts all have in common? They all **oscillate**—-that is, they move back and forth between two points. Many systems oscillate, and they have certain characteristics in common. All oscillations involve force and energy. You push a child in a swing to get the motion started. The energy of atoms vibrating in a crystal can be increased with heat. You put energy into a guitar string when you pluck it.

Some oscillations create **waves**. A guitar creates sound waves. You can make water waves in a swimming pool by slapping the water with your hand. You can no doubt think of other types of waves. Some, such as water waves, are visible. Some, such as sound waves, are not. But *every wave is a disturbance that moves from its source and carries energy*. Other examples of waves include earthquakes and visible light. Even subatomic particles, such as electrons, can behave like waves.

By studying oscillatory motion and waves, we shall find that a small number of underlying principles describe all of them and that wave phenomena are more common than you have ever imagined. We begin by studying the type of force that underlies the simplest oscillations and waves. We will then expand our exploration of oscillatory motion and waves to

include concepts such as simple harmonic motion, uniform circular motion, and damped harmonic motion. Finally, we will explore what happens when two or more waves share the same space, in the phenomena known as superposition and interference.

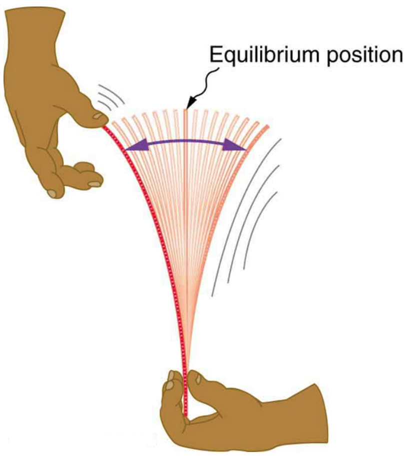
Glossary

oscillate
: moving back and forth regularly between two points

wave
: a disturbance that moves from its source and carries energy

Hooke’s Law: Stress and Strain Revisited

- Explain Newton’s third law of motion with respect to stress and deformation.
- Describe the restoration of force and displacement.
- Calculate the energy in Hooke’s Law of deformation, and the stored energy in a spring.



When displaced from its vertical equilibrium position, this plastic ruler oscillates back and forth because of the restoring force opposing displacement. When the ruler is on the left, there is a force to the right, and vice versa.

Newton’s first law implies that an object oscillating back and forth is experiencing forces. Without force, the object would move in a straight line

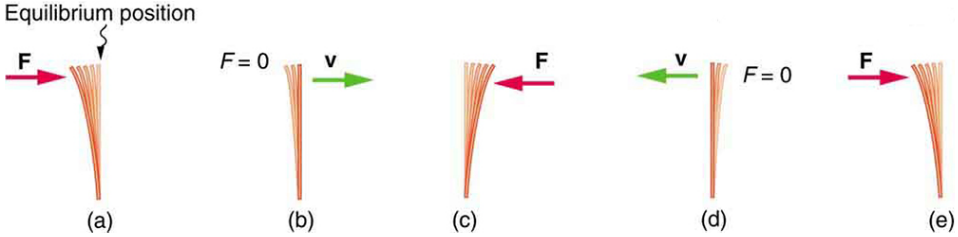
at a constant speed rather than oscillate. Consider, for example, plucking a plastic ruler to the left as shown in [link]. The deformation of the ruler creates a force in the opposite direction, known as a **restoring force**. Once released, the restoring force causes the ruler to move back toward its stable equilibrium position, where the net force on it is zero. However, by the time the ruler gets there, it gains momentum and continues to move to the right, producing the opposite deformation. It is then forced to the left, back through equilibrium, and the process is repeated until dissipative forces dampen the motion. These forces remove mechanical energy from the system, gradually reducing the motion until the ruler comes to rest.

The simplest oscillations occur when the restoring force is directly proportional to displacement. When stress and strain were covered in Newton's Third Law of Motion, the name was given to this relationship between force and displacement was Hooke's law:

Equation:

$$F = -\mathrm{kx}.$$

Here, F is the restoring force, x is the displacement from equilibrium or **deformation**, and k is a constant related to the difficulty in deforming the system. The minus sign indicates the restoring force is in the direction opposite to the displacement.

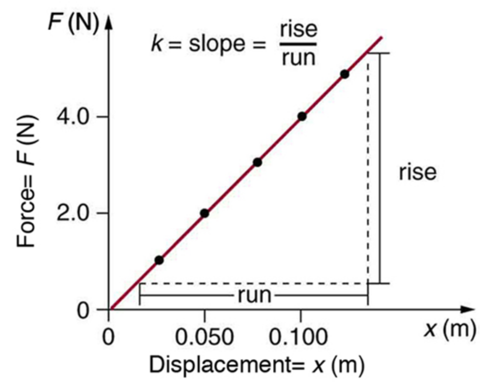


(a) The plastic ruler has been released, and the restoring force is returning the ruler to its equilibrium position. (b) The net force is zero at the equilibrium position, but the ruler has momentum and continues to move to the right. (c) The restoring force is in the opposite direction. It stops the ruler and moves it back toward equilibrium again. (d) Now the ruler has momentum to the left. (e) In the absence of damping

(caused by frictional forces), the ruler reaches its original position. From there, the motion will repeat itself.

The **force constant** k is related to the rigidity (or stiffness) of a system—the larger the force constant, the greater the restoring force, and the stiffer the system. The units of k are newtons per meter (N/m). For example, k is directly related to Young's modulus when we stretch a string. [link] shows a graph of the absolute value of the restoring force versus the displacement for a system that can be described by Hooke's law—a simple spring in this case. The slope of the graph equals the force constant k in newtons per meter. A common physics laboratory exercise is to measure restoring forces created by springs, determine if they follow Hooke's law, and calculate their force constants if they do.

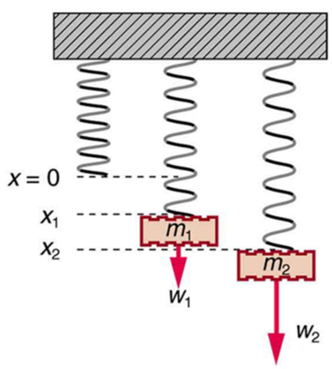
a)



m (kg)	w (N)	x (m)
0.000	0.00	0.000
0.100	0.98	0.025
0.200	1.96	0.050
0.300	2.94	0.076
0.400	3.92	0.099
0.500	4.90	0.127

b)

m (kg)	w (N)	x (m)
0.000	0.00	0.000
0.100	0.98	0.025
0.200	1.96	0.050
0.300	2.94	0.076
0.400	3.92	0.099
0.500	4.90	0.127



(a) A graph of absolute value of the restoring force versus displacement is

displayed. The fact that the graph is a straight line means that the system obeys Hooke's law. The slope of the graph is the force constant k. (b) The data in the graph were generated by measuring the displacement of a spring from equilibrium while supporting various weights. The restoring force equals the weight supported, if the mass is stationary.

Example:
How Stiff Are Car Springs?

The mass of a car increases due to the introduction of a passenger. This affects the displacement of

the car on its suspension system. (credit: exfordy on Flickr)

What is the force constant for the suspension system of a car that settles 1.20 cm when an 80.0-kg person gets in?

Strategy

Consider the car to be in its equilibrium position $x = 0$ before the person gets in. The car then settles down 1.20 cm, which means it is displaced to a position $x = -1.20 \times 10^{-2}$ m. At that point, the springs supply a restoring force F equal to the person's weight $w = \text{mg} = (80.0 \text{ kg})\left(9.80 \text{ m/s}^2\right) = 784 \text{ N}$. We take this force to be F in Hooke's law. Knowing F and x, we can then solve the force constant k.

Solution

1. Solve Hooke's law, $F = -\text{kx}$, for k:
   Equation:
   $$k = -\frac{F}{x}.$$
   Substitute known values and solve k:
   Equation:
   $$\begin{aligned} k &= -\frac{784 \text{ N}}{-1.20 \times 10^{-2} \text{ m}} \\ &= 6.53 \times 10^{4} \text{ N/m}. \end{aligned}$$

Discussion

Note that F and x have opposite signs because they are in opposite directions—the restoring force is up, and the displacement is down. Also, note that the car would oscillate up and down when the person got in if it

were not for damping (due to frictional forces) provided by shock absorbers. Bouncing cars are a sure sign of bad shock absorbers.

Energy in Hooke's Law of Deformation

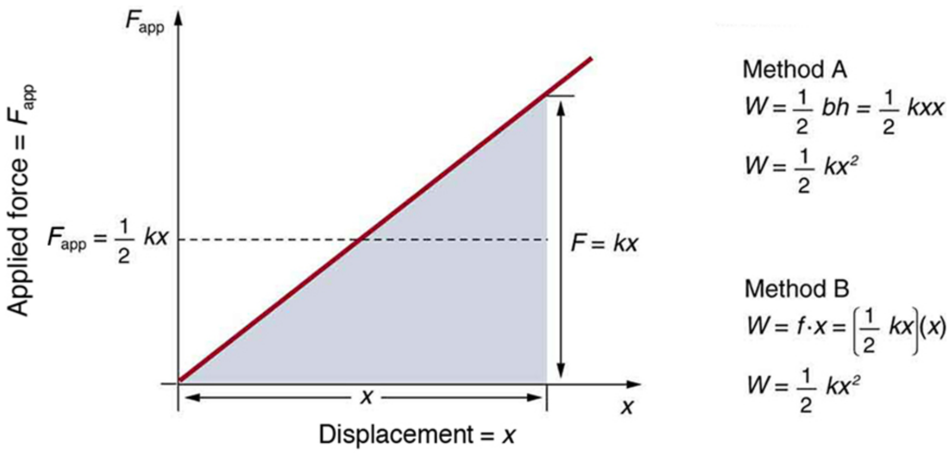
In order to produce a deformation, work must be done. That is, a force must be exerted through a distance, whether you pluck a guitar string or compress a car spring. If the only result is deformation, and no work goes into thermal, sound, or kinetic energy, then all the work is initially stored in the deformed object as some form of potential energy. The potential energy stored in a spring is $\mathrm{PE_{el}} = \frac{1}{2}\mathrm{kx}^2$. Here, we generalize the idea to elastic potential energy for a deformation of any system that can be described by Hooke's law. Hence,

Equation:

$$\mathrm{PE_{el}} = \frac{1}{2}\mathrm{kx}^2,$$

where $\mathrm{PE_{el}}$ is the **elastic potential energy** stored in any deformed system that obeys Hooke's law and has a displacement x from equilibrium and a force constant k.

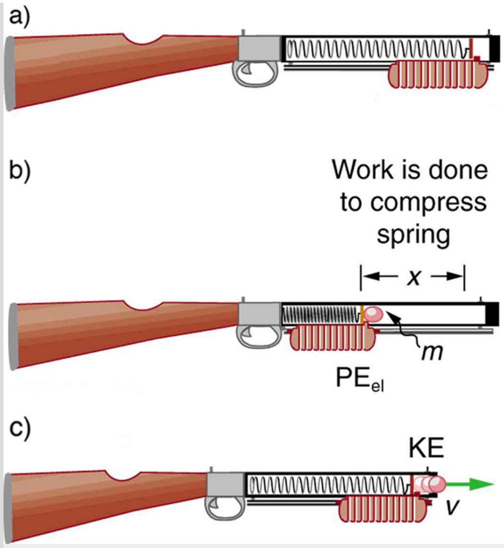
It is possible to find the work done in deforming a system in order to find the energy stored. This work is performed by an applied force F_{app}. The applied force is exactly opposite to the restoring force (action-reaction), and so $F_{\text{app}} = \mathrm{kx}$. [link] shows a graph of the applied force versus deformation x for a system that can be described by Hooke's law. Work done on the system is force multiplied by distance, which equals the area under the curve or $(1/2)\mathrm{kx}^2$(Method A in the figure). Another way to determine the work is to note that the force increases linearly from 0 to kx, so that the average force is $(1/2)\,\mathrm{kx}$, the distance moved is x, and thus $W = F_{\text{app}}d = [(1/2)\mathrm{kx}](x) = (1/2)\mathrm{kx}^2$ (Method B in the figure).



A graph of applied force versus distance for the deformation of a system that can be described by Hooke's law is displayed. The work done on the system equals the area under the graph or the area of the triangle, which is half its base multiplied by its height, or $W = (1/2)\mathrm{kx}^2$.

Example:
Calculating Stored Energy: A Tranquilizer Gun Spring

We can use a toy gun's spring mechanism to ask and answer two simple questions: (a) How much energy is stored in the spring of a tranquilizer gun that has a force constant of 50.0 N/m and is compressed 0.150 m? (b) If you neglect friction and the mass of the spring, at what speed will a 2.00-g projectile be ejected from the gun?



(a) In this image of the gun, the spring is uncompressed before being cocked. (b) The spring has been compressed a distance x, and the projectile is in place. (c) When released, the spring converts elastic potential energy $\mathrm{PE_{el}}$ into kinetic energy.

Strategy for a

(a): The energy stored in the spring can be found directly from elastic potential energy equation, because k and x are given.

Solution for a

Entering the given values for k and x yields

Equation:

$$\begin{aligned} \mathrm{PE_{el}} &= \frac{1}{2}\mathrm{kx}^2 = \frac{1}{2}(50.0\ \mathrm{N/m})(0.150\ \mathrm{m})^2 = 0.563\ \mathrm{N}\cdot\mathrm{m} \\ &= 0.563\ \mathrm{J} \end{aligned}$$

Strategy for b

Because there is no friction, the potential energy is converted entirely into kinetic energy. The expression for kinetic energy can be solved for the projectile's speed.

Solution for b

1. Identify known quantities:
   Equation:

$$\mathrm{KE_f} = \mathrm{PE_{el}} \;\text{ or }\; 1/2mv^2 = (1/2)kx^2 = \mathrm{PE_{el}} = 0.563\ \mathrm{J}$$

2. Solve for v:
   Equation:

$$v = \left[\frac{2\mathrm{PE_{el}}}{m}\right]^{1/2} = \left[\frac{2(0.563\ \mathrm{J})}{0.002\ \mathrm{kg}}\right]^{1/2} = 23.7(\mathrm{J/kg})^{1/2}$$

3. Convert units: $23.7\ \mathrm{m/s}$

Discussion

(a) and (b): This projectile speed is impressive for a tranquilizer gun (more than 80 km/h). The numbers in this problem seem reasonable. The force needed to compress the spring is small enough for an adult to manage, and the energy imparted to the dart is small enough to limit the damage it might do. Yet, the speed of the dart is great enough for it to travel an acceptable distance.

Exercise:
Check your Understanding

Problem:

Envision holding the end of a ruler with one hand and deforming it with the other. When you let go, you can see the oscillations of the ruler. In what way could you modify this simple experiment to increase the rigidity of the system?

Solution:
Answer

You could hold the ruler at its midpoint so that the part of the ruler that oscillates is half as long as in the original experiment.

Exercise:
Check your Understanding

Problem:

If you apply a deforming force on an object and let it come to equilibrium, what happened to the work you did on the system?

Solution:
Answer

It was stored in the object as potential energy.

Section Summary

- An oscillation is a back and forth motion of an object between two points of deformation.
- An oscillation may create a wave, which is a disturbance that propagates from where it was created.
- The simplest type of oscillations and waves are related to systems that can be described by Hooke's law:
  Equation:
  $$F = -\mathrm{kx},$$
  where F is the restoring force, x is the displacement from equilibrium or deformation, and k is the force constant of the system.
- Elastic potential energy $\mathrm{PE_{el}}$ stored in the deformation of a system that can be described by Hooke's law is given by
  Equation:

$$PE_{el} = (1/2)kx^2.$$

Conceptual Questions

Exercise:

Problem:

Describe a system in which elastic potential energy is stored.

Problems & Exercises

Exercise:

Problem:

Fish are hung on a spring scale to determine their mass (most fishermen feel no obligation to truthfully report the mass).

(a) What is the force constant of the spring in such a scale if it the spring stretches 8.00 cm for a 10.0 kg load?

(b) What is the mass of a fish that stretches the spring 5.50 cm?

(c) How far apart are the half-kilogram marks on the scale?

Solution:

(a) $1.23 \times 10^3 \text{ N/m}$

(b) 6.88 kg

(c) 4.00 mm

Exercise:

Problem:

It is weigh-in time for the local under-85-kg rugby team. The bathroom scale used to assess eligibility can be described by Hooke's law and is depressed 0.75 cm by its maximum load of 120 kg. (a) What is the spring's effective spring constant? (b) A player stands on the scales and depresses it by 0.48 cm. Is he eligible to play on this under-85 kg team?

Exercise:

Problem:

One type of BB gun uses a spring-driven plunger to blow the BB from its barrel. (a) Calculate the force constant of its plunger's spring if you must compress it 0.150 m to drive the 0.0500-kg plunger to a top speed of 20.0 m/s. (b) What force must be exerted to compress the spring?

Solution:

(a) 889 N/m

(b) 133 N

Exercise:

Problem:

(a) The springs of a pickup truck act like a single spring with a force constant of $1.30 \times 10^5 \text{ N/m}$. By how much will the truck be depressed by its maximum load of 1000 kg?

(b) If the pickup truck has four identical springs, what is the force constant of each?

Exercise:

Problem:

When an 80.0-kg man stands on a pogo stick, the spring is compressed 0.120 m.

(a) What is the force constant of the spring? (b) Will the spring be compressed more when he hops down the road?

Solution:

(a) $6.53 \times 10^3 \text{ N/m}$

(b) Yes

Exercise:

Problem:

A spring has a length of 0.200 m when a 0.300-kg mass hangs from it, and a length of 0.750 m when a 1.95-kg mass hangs from it. (a) What is the force constant of the spring? (b) What is the unloaded length of the spring?

Glossary

deformation
: displacement from equilibrium

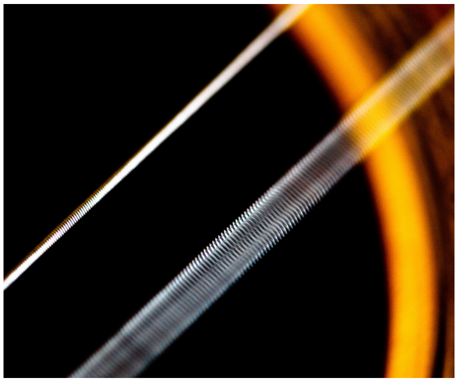
elastic potential energy
: potential energy stored as a result of deformation of an elastic object, such as the stretching of a spring

force constant
: a constant related to the rigidity of a system: the larger the force constant, the more rigid the system; the force constant is represented by *k*

restoring force
: force acting in opposition to the force caused by a deformation

Period and Frequency in Oscillations

- Observe the vibrations of a guitar string.
- Determine the frequency of oscillations.



The strings on this guitar vibrate at regular time intervals. (credit: JAR)

When you pluck a guitar string, the resulting sound has a steady tone and lasts a long time. Each successive vibration of the string takes the same time as the previous one. We define **periodic motion** to be a motion that repeats itself at regular time intervals, such as exhibited by the guitar string or by an object on a spring moving up and down. The time to complete one oscillation remains constant and is called the **period** T. Its units are usually seconds, but may be any convenient unit of time. The word period refers to the time for some event whether repetitive or not; but we shall be primarily interested in periodic motion, which is by definition repetitive. A concept closely related to period is the frequency of an event. For example, if you get a paycheck twice a month, the frequency of payment is two per month and the period between checks is half a month. **Frequency** f is defined to be the number of events per unit time. For periodic motion, frequency is the number of oscillations per unit time. The relationship between frequency and period is

Equation:

$$f = \frac{1}{T}.$$

The SI unit for frequency is the *cycle per second*, which is defined to be a *hertz* (Hz):
Equation:

$$1\text{ Hz} = 1\frac{\text{cycle}}{\text{sec}}\text{ or }1\text{ Hz} = \frac{1}{\text{s}}$$

A cycle is one complete oscillation. Note that a vibration can be a single or multiple event, whereas oscillations are usually repetitive for a significant number of cycles.

Example:
Determine the Frequency of Two Oscillations: Medical Ultrasound and the Period of Middle C
We can use the formulas presented in this module to determine both the frequency based on known oscillations and the oscillation based on a known frequency. Let's try one example of each. (a) A medical imaging device produces ultrasound by oscillating with a period of 0.400 µs. What is the frequency of this oscillation? (b) The frequency of middle C on a typical musical instrument is 264 Hz. What is the time for one complete oscillation?
Strategy
Both questions (a) and (b) can be answered using the relationship between period and frequency. In question (a), the period T is given and we are asked to find frequency f. In question (b), the frequency f is given and we are asked to find the period T.
Solution a

1. Substitute 0.400 µs for T in $f = \frac{1}{T}$:
   Equation:

$$f = \frac{1}{T} = \frac{1}{0.400 \times 10^{-6}\ \text{s}}.$$

Solve to find
Equation:

$$f = 2.50 \times 10^{6}\ \text{Hz}.$$

Discussion a
The frequency of sound found in (a) is much higher than the highest frequency that humans can hear and, therefore, is called ultrasound. Appropriate oscillations at this frequency generate ultrasound used for noninvasive medical diagnoses, such as observations of a fetus in the womb.
Solution b

1. Identify the known values:

   The time for one complete oscillation is the period T:
   Equation:

   $$f = \frac{1}{T}.$$

2. Solve for T:
   Equation:

   $$T = \frac{1}{f}.$$

3. Substitute the given value for the frequency into the resulting expression:
   Equation:

   $$T = \frac{1}{f} = \frac{1}{264\ \text{Hz}} = \frac{1}{264\ \text{cycles/s}} = 3.79 \times 10^{-3}\ \text{s} = 3.79\ \text{ms}.$$

Discussion
The period found in (b) is the time per cycle, but this value is often quoted as simply the time in convenient units (ms or milliseconds in this case).

Exercise:
Check your Understanding

Problem:

Identify an event in your life (such as receiving a paycheck) that occurs regularly. Identify both the period and frequency of this event.

Solution:

I visit my parents for dinner every other Sunday. The frequency of my visits is 26 per calendar year. The period is two weeks.

Section Summary

- Periodic motion is a repetitious oscillation.
- The time for one oscillation is the period T.
- The number of oscillations per unit time is the frequency f.
- These quantities are related by
  Equation:

$$f = \frac{1}{T}.$$

Problems & Exercises

Exercise:

Problem: What is the period of 60.0 Hz electrical power?

Solution:

16.7 ms

Exercise:

Problem:

If your heart rate is 150 beats per minute during strenuous exercise, what is the time per beat in units of seconds?

Solution:

0.400 s/beats

Exercise:

Problem:

Find the frequency of a tuning fork that takes 2.50×10^{-3} s to complete one oscillation.

Solution:

400 Hz

Exercise:

Problem:

A stroboscope is set to flash every 8.00×10^{-5} s. What is the frequency of the flashes?

Solution:

12,500 Hz

Exercise:

Problem:

A tire has a tread pattern with a crevice every 2.00 cm. Each crevice makes a single vibration as the tire moves. What is the frequency of these vibrations if the car moves at 30.0 m/s?

Solution:

1.50 kHz

Exercise:

Problem: **Engineering Application**

Each piston of an engine makes a sharp sound every other revolution of the engine. (a) How fast is a race car going if its eight-cylinder engine emits a sound of frequency 750 Hz, given that the engine makes 2000 revolutions per kilometer? (b) At how many revolutions per minute is the engine rotating?

Solution:

(a) 93.8 m/s

(b) $11.3 \times 10^3 \text{ rev/min}$

Glossary

period
: time it takes to complete one oscillation

periodic motion
: motion that repeats itself at regular time intervals

frequency
: number of events per unit of time

Simple Harmonic Motion: A Special Periodic Motion

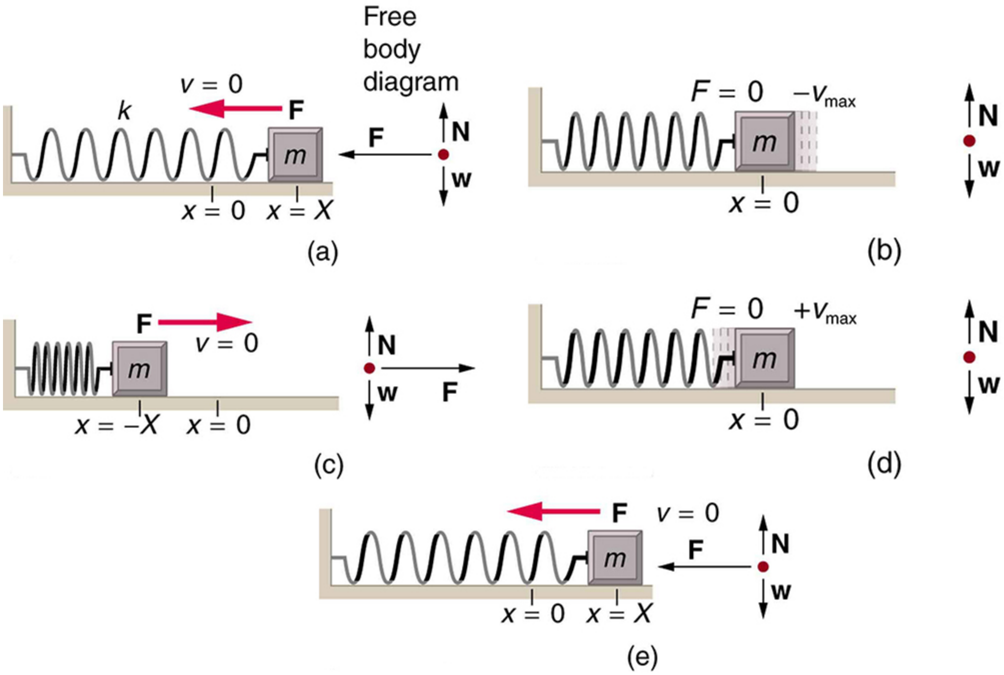
- Describe a simple harmonic oscillator.
- Explain the link between simple harmonic motion and waves.

The oscillations of a system in which the net force can be described by Hooke’s law are of special importance, because they are very common. They are also the simplest oscillatory systems. **Simple Harmonic Motion** (SHM) is the name given to oscillatory motion for a system where the net force can be described by Hooke’s law, and such a system is called a **simple harmonic oscillator**. If the net force can be described by Hooke’s law and there is no *damping* (by friction or other non-conservative forces), then a simple harmonic oscillator will oscillate with equal displacement on either side of the equilibrium position, as shown for an object on a spring in [link]. The maximum displacement from equilibrium is called the **amplitude** X. The units for amplitude and displacement are the same, but depend on the type of oscillation. For the object on the spring, the units of amplitude and displacement are meters; whereas for sound oscillations, they have units of pressure (and other types of oscillations have yet other units). Because amplitude is the maximum displacement, it is related to the energy in the oscillation.

Note:

Take-Home Experiment: SHM and the Marble

Find a bowl or basin that is shaped like a hemisphere on the inside. Place a marble inside the bowl and tilt the bowl periodically so the marble rolls from the bottom of the bowl to equally high points on the sides of the bowl. Get a feel for the force required to maintain this periodic motion. What is the restoring force and what role does the force you apply play in the simple harmonic motion (SHM) of the marble?



An object attached to a spring sliding on a frictionless surface is an uncomplicated simple harmonic oscillator. When displaced from equilibrium, the object performs simple harmonic motion that has an amplitude X and a period T. The object's maximum speed occurs as it passes through equilibrium. The stiffer the spring is, the smaller the period T. The greater the mass of the object is, the greater the period T.

What is so significant about simple harmonic motion? One special thing is that the period T and frequency f of a simple harmonic oscillator are independent of amplitude. The string of a guitar, for example, will oscillate with the same frequency whether plucked gently or hard. Because the period is constant, a simple harmonic oscillator can be used as a clock.

Two important factors do affect the period of a simple harmonic oscillator. The period is related to how stiff the system is. A very stiff object has a large force constant k, which causes the system to have a smaller period. For example, you can adjust a diving board's stiffness—the stiffer it is, the

faster it vibrates, and the shorter its period. Period also depends on the mass of the oscillating system. The more massive the system is, the longer the period. For example, a heavy person on a diving board bounces up and down more slowly than a light one.

In fact, the mass m and the force constant k are the *only* factors that affect the period and frequency of simple harmonic motion.

Note:
Period of Simple Harmonic Oscillator
The *period of a simple harmonic oscillator* is given by
Equation:

$$T = 2\pi\sqrt{\frac{m}{k}}$$

and, because $f = 1/T$, the *frequency of a simple harmonic oscillator* is
Equation:

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}.$$

Note that neither T nor f has any dependence on amplitude.

Note:
Take-Home Experiment: Mass and Ruler Oscillations
Find two identical wooden or plastic rulers. Tape one end of each ruler firmly to the edge of a table so that the length of each ruler that protrudes from the table is the same. On the free end of one ruler tape a heavy object such as a few large coins. Pluck the ends of the rulers at the same time and observe which one undergoes more cycles in a time period, and measure the period of oscillation of each of the rulers.

Example:
Calculate the Frequency and Period of Oscillations: Bad Shock Absorbers in a Car

If the shock absorbers in a car go bad, then the car will oscillate at the least provocation, such as when going over bumps in the road and after stopping (See [link]). Calculate the frequency and period of these oscillations for such a car if the car's mass (including its load) is 900 kg and the force constant (k) of the suspension system is $6.53 \times 10^4\ \text{N/m}$.

Strategy

The frequency of the car's oscillations will be that of a simple harmonic oscillator as given in the equation $f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$. The mass and the force constant are both given.

Solution

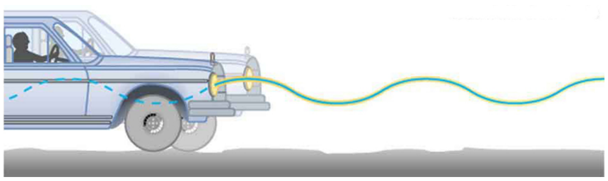
1. Enter the known values of *k* and *m*:
   Equation:
   $$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{6.53 \times 10^4\ \text{N/m}}{900\ \text{kg}}}.$$
2. Calculate the frequency:
   Equation:
   $$\frac{1}{2\pi}\sqrt{72.6/\text{s}^{-2}} = 1.3656/\text{s}^{-1} \approx 1.36/\text{s}^{-1} = 1.36\ \text{Hz}.$$
3. You could use $T = 2\pi\sqrt{\frac{m}{k}}$ to calculate the period, but it is simpler to use the relationship $T = 1/f$ and substitute the value just found for f:
   Equation:
   $$T = \frac{1}{f} = \frac{1}{1.356\ \text{Hz}} = 0.738\ \text{s}.$$

Discussion

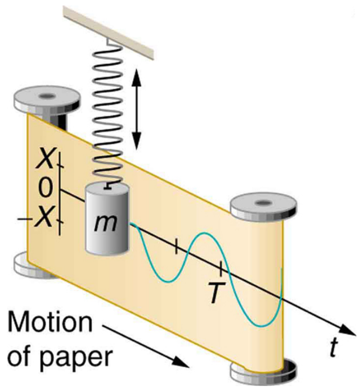
The values of T and f both seem about right for a bouncing car. You can observe these oscillations if you push down hard on the end of a car and let go.

The Link between Simple Harmonic Motion and Waves

If a time-exposure photograph of the bouncing car were taken as it drove by, the headlight would make a wavelike streak, as shown in [link]. Similarly, [link] shows an object bouncing on a spring as it leaves a wavelike "trace of its position on a moving strip of paper. Both waves are sine functions. All simple harmonic motion is intimately related to sine and cosine waves.



The bouncing car makes a wavelike motion. If the restoring force in the suspension system can be described only by Hooke's law, then the wave is a sine function. (The wave is the trace produced by the headlight as the car moves to the right.)



The vertical position of an object bouncing on a spring is recorded on a strip of moving paper, leaving a sine wave.

The displacement as a function of time *t* in any simple harmonic motion—that is, one in which the net restoring force can be described by Hooke's law, is given by

Equation:

$$x(t) = X\cos\frac{2\pi t}{T},$$

where X is amplitude. At $t = 0$, the initial position is $x_0 = X$, and the displacement oscillates back and forth with a period T. (When $t = T$, we get $x = X$ again because cos $2\pi = 1$.). Furthermore, from this expression for x, the velocity v as a function of time is given by:

Equation:

$$v(t) = -v_{\text{max}}\sin\left(\frac{2\pi t}{T}\right),$$

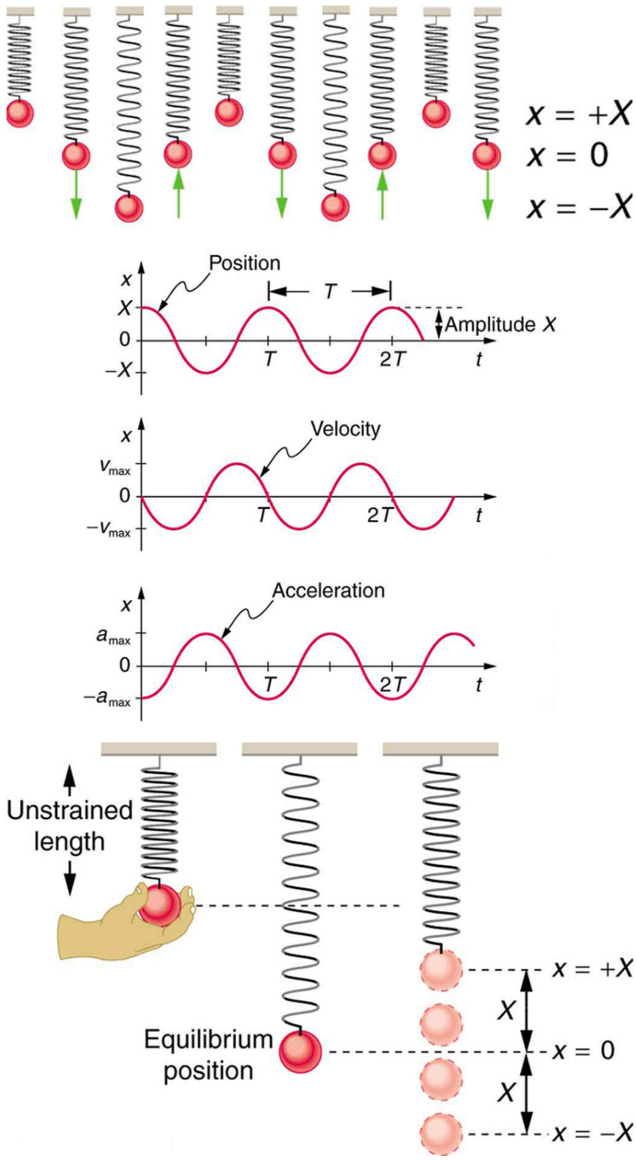
where $v_{\text{max}} = 2\pi X/T = X\sqrt{k/m}$. The object has zero velocity at maximum displacement—for example, $v = 0$ when $t = 0$, and at that time $x = X$. The minus sign in the first equation for $v(t)$ gives the correct direction for the velocity. Just after the start of the motion, for instance, the velocity is negative because the system is moving back toward the equilibrium point. Finally, we can get an expression for acceleration using Newton's second law. [Then we have $x(t)$, $v(t)$, t, and $a(t)$, the quantities needed for kinematics and a description of simple harmonic motion.] According to Newton's second law, the acceleration is $a = F/m = \text{kx}/m$. So, $a(t)$ is also a cosine function:

Equation:

$$a(t) = -\frac{\text{kX}}{m}\cos\frac{2\pi t}{T}.$$

Hence, $a(t)$ is directly proportional to and in the opposite direction to $x(t)$.

[link] shows the simple harmonic motion of an object on a spring and presents graphs of $x(t)$,$v(t\ )$, and $a(t)$ versus time.



Graphs of $x(t)$, $v(t)$, and $a(t)$ versus t for the motion of an object on a spring. The net force on the object can be described by Hooke's law, and so the object undergoes simple harmonic motion. Note that the initial position has the vertical displacement at its maximum value X; v is initially zero and then negative as the object moves down; and the initial acceleration

is negative, back toward the equilibrium position and becomes zero at that point.

The most important point here is that these equations are mathematically straightforward and are valid for all simple harmonic motion. They are very useful in visualizing waves associated with simple harmonic motion, including visualizing how waves add with one another.

Exercise:
Check Your Understanding

Problem:

Suppose you pluck a banjo string. You hear a single note that starts out loud and slowly quiets over time. Describe what happens to the sound waves in terms of period, frequency and amplitude as the sound decreases in volume.

Solution:

Frequency and period remain essentially unchanged. Only amplitude decreases as volume decreases.

Exercise:
Check Your Understanding

Problem:

A babysitter is pushing a child on a swing. At the point where the swing reaches x, where would the corresponding point on a wave of this motion be located?

Solution:

x is the maximum deformation, which corresponds to the amplitude of the wave. The point on the wave would either be at the very top or the very bottom of the curve.

Note:
PhET Explorations: Masses and Springs
A realistic mass and spring laboratory. Hang masses from springs and adjust the spring stiffness and damping. You can even slow time. Transport the lab to different planets. A chart shows the kinetic, potential, and thermal energy for each spring.
https://phet.colorado.edu/sims/mass-spring-lab/mass-spring-lab_en.html

Section Summary

- Simple harmonic motion is oscillatory motion for a system that can be described only by Hooke’s law. Such a system is also called a simple harmonic oscillator.
- Maximum displacement is the amplitude X. The period T and frequency f of a simple harmonic oscillator are given by

  $T = 2\pi\sqrt{\frac{m}{k}}$ and $f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$, where m is the mass of the system.
- Displacement in simple harmonic motion as a function of time is given by $x(t) = X\cos\frac{2\pi t}{T}$.
- The velocity is given by $v(t) = -v_{\text{max}}\sin\frac{2\pi t}{T}$, where $v_{\text{max}} = \sqrt{k/m}X$.
- The acceleration is found to be $a(t) = -\frac{kX}{m}\cos\frac{2\pi t}{T}$.

Conceptual Questions

Exercise:

Problem:

What conditions must be met to produce simple harmonic motion?

Exercise:

Problem:

(a) If frequency is not constant for some oscillation, can the oscillation be simple harmonic motion?

(b) Can you think of any examples of harmonic motion where the frequency may depend on the amplitude?

Exercise:

Problem:

Give an example of a simple harmonic oscillator, specifically noting how its frequency is independent of amplitude.

Exercise:

Problem:

Explain why you expect an object made of a stiff material to vibrate at a higher frequency than a similar object made of a spongy material.

Exercise:

Problem:

As you pass a freight truck with a trailer on a highway, you notice that its trailer is bouncing up and down slowly. Is it more likely that the trailer is heavily loaded or nearly empty? Explain your answer.

Exercise:

Problem:

Some people modify cars to be much closer to the ground than when manufactured. Should they install stiffer springs? Explain your answer.

Problems & Exercises

Exercise:

Problem:

A type of cuckoo clock keeps time by having a mass bouncing on a spring, usually something cute like a cherub in a chair. What force constant is needed to produce a period of 0.500 s for a 0.0150-kg mass?

Solution:

$2.37\ \mathrm{N/m}$

Exercise:

Problem:

If the spring constant of a simple harmonic oscillator is doubled, by what factor will the mass of the system need to change in order for the frequency of the motion to remain the same?

Exercise:

Problem:

A 0.500-kg mass suspended from a spring oscillates with a period of 1.50 s. How much mass must be added to the object to change the period to 2.00 s?

Solution:

0.389 kg

Exercise:

Problem:

By how much leeway (both percentage and mass) would you have in the selection of the mass of the object in the previous problem if you did not wish the new period to be greater than 2.01 s or less than 1.99 s?

Exercise:

Problem:

Suppose you attach the object with mass m to a vertical spring originally at rest, and let it bounce up and down. You release the object from rest at the spring's original rest length. (a) Show that the spring exerts an upward force of $2.00\ \text{mg}$ on the object at its lowest point. (b) If the spring has a force constant of $10.0\ \text{N/m}$ and a 0.25-kg-mass object is set in motion as described, find the amplitude of the oscillations. (c) Find the maximum velocity.

Exercise:

Problem:

A diver on a diving board is undergoing simple harmonic motion. Her mass is 55.0 kg and the period of her motion is 0.800 s. The next diver is a male whose period of simple harmonic oscillation is 1.05 s. What is his mass if the mass of the board is negligible?

Solution:

94.7 kg

Exercise:

Problem:

Suppose a diving board with no one on it bounces up and down in a simple harmonic motion with a frequency of 4.00 Hz. The board has an effective mass of 10.0 kg. What is the frequency of the simple harmonic motion of a 75.0-kg diver on the board?

Exercise:

Problem:

This child's toy relies on springs to keep infants entertained. (credit: By Humboldthead, Flickr)

The device pictured in [link] entertains infants while keeping them from wandering. The child bounces in a harness suspended from a door frame by a spring constant.

(a) If the spring stretches 0.250 m while supporting an 8.0-kg child, what is its spring constant?

(b) What is the time for one complete bounce of this child? (c) What is the child's maximum velocity if the amplitude of her bounce is 0.200 m?

Exercise:

Problem:

A 90.0-kg skydiver hanging from a parachute bounces up and down with a period of 1.50 s. What is the new period of oscillation when a second skydiver, whose mass is 60.0 kg, hangs from the legs of the first, as seen in [link].

The oscillations of one skydiver are about to be affected by a second skydiver. (credit: U.S. Army, www.army.mil)

Solution:

1.94 s

Glossary

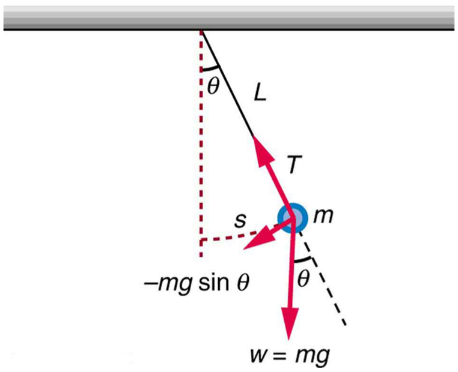
amplitude
: the maximum displacement from the equilibrium position of an object oscillating around the equilibrium position

simple harmonic motion
: the oscillatory motion in a system where the net force can be described by Hooke’s law

simple harmonic oscillator
: a device that implements Hooke’s law, such as a mass that is attached to a spring, with the other end of the spring being connected to a rigid support such as a wall

The Simple Pendulum

- Measure acceleration due to gravity.



A simple pendulum has a small-diameter bob and a string that has a very small mass but is strong enough not to stretch appreciably. The linear displacement from equilibrium is s, the length of the arc. Also shown are the forces on the bob, which result in a net force of $-\text{mg}\ \text{sin}\theta$ toward the equilibrium position—that is, a restoring force.

Pendulums are in common usage. Some have crucial uses, such as in clocks; some are for fun, such as a child’s swing; and some are just there, such as the sinker on a fishing line. For small displacements, a pendulum is a simple harmonic oscillator. A **simple pendulum** is defined to have an

object that has a small mass, also known as the pendulum bob, which is suspended from a light wire or string, such as shown in [link]. Exploring the simple pendulum a bit further, we can discover the conditions under which it performs simple harmonic motion, and we can derive an interesting expression for its period.

We begin by defining the displacement to be the arc length s. We see from [link] that the net force on the bob is tangent to the arc and equals $-\text{mg sin } \theta$. (The weight mg has components $\text{mg cos } \theta$ along the string and $\text{mg sin } \theta$ tangent to the arc.) Tension in the string exactly cancels the component $\text{mg cos } \theta$ parallel to the string. This leaves a *net* restoring force back toward the equilibrium position at $\theta = 0$.

Now, if we can show that the restoring force is directly proportional to the displacement, then we have a simple harmonic oscillator. In trying to determine if we have a simple harmonic oscillator, we should note that for small angles (less than about 15^{o}), $\text{sin } \theta \approx \theta$ ($\text{sin } \theta$ and θ differ by about 1% or less at smaller angles). Thus, for angles less than about 15^{o}, the restoring force F is

Equation:

$$F \approx -\text{mg}\theta.$$

The displacement s is directly proportional to θ. When θ is expressed in radians, the arc length in a circle is related to its radius (L in this instance) by:

Equation:

$$s = L\theta,$$

so that

Equation:

$$\theta = \frac{s}{L}.$$

For small angles, then, the expression for the restoring force is:
Equation:

$$F \approx -\frac{\mathrm{mg}}{L} s$$

This expression is of the form:
Equation:

$$F = -\mathrm{kx},$$

where the force constant is given by $k = mg/L$ and the displacement is given by $x = s$. For angles less than about 15º, the restoring force is directly proportional to the displacement, and the simple pendulum is a simple harmonic oscillator.

Using this equation, we can find the period of a pendulum for amplitudes less than about 15º. For the simple pendulum:
Equation:

$$T = 2\pi \sqrt{\frac{m}{k}} = 2\pi \sqrt{\frac{m}{\mathrm{mg}/L}}$$

Thus,
Equation:

$$T = 2\pi \sqrt{\frac{L}{g}}$$

for the period of a simple pendulum. This result is interesting because of its simplicity. The only things that affect the period of a simple pendulum are its length and the acceleration due to gravity. The period is completely independent of other factors, such as mass. As with simple harmonic oscillators, the period T for a pendulum is nearly independent of amplitude,

especially if θ is less than about 15º. Even simple pendulum clocks can be finely adjusted and accurate.

Note the dependence of T on g. If the length of a pendulum is precisely known, it can actually be used to measure the acceleration due to gravity. Consider the following example.

Example:
Measuring Acceleration due to Gravity: The Period of a Pendulum
What is the acceleration due to gravity in a region where a simple pendulum having a length 75.000 cm has a period of 1.7357 s?
Strategy
We are asked to find g given the period T and the length L of a pendulum. We can solve $T = 2\pi\sqrt{\frac{L}{g}}$ for g, assuming only that the angle of deflection is less than 15º.
Solution

1. Square $T = 2\pi\sqrt{\frac{L}{g}}$ and solve for g:
   Equation:
   $$g = 4\pi^2\frac{L}{T^2}.$$
2. Substitute known values into the new equation:
   Equation:
   $$g = 4\pi^2\frac{0.75000\text{ m}}{(1.7357\text{ s})^2}.$$
3. Calculate to find g:
   Equation:
   $$g = 9.8281\text{ m/s}^2.$$

Discussion

This method for determining *g* can be very accurate. This is why length and period are given to five digits in this example. For the precision of the approximation $\sin\theta \approx \theta$ to be better than the precision of the pendulum length and period, the maximum displacement angle should be kept below about 0.5°.

Note:

Making Career Connections

Knowing *g* can be important in geological exploration; for example, a map of *g* over large geographical regions aids the study of plate tectonics and helps in the search for oil fields and large mineral deposits.

Note:

Take Home Experiment: Determining *g*

Use a simple pendulum to determine the acceleration due to gravity *g* in your own locale. Cut a piece of a string or dental floss so that it is about 1 m long. Attach a small object of high density to the end of the string (for example, a metal nut or a car key). Starting at an angle of less than 10°, allow the pendulum to swing and measure the pendulum's period for 10 oscillations using a stopwatch. Calculate *g*. How accurate is this measurement? How might it be improved?

Exercise:

Check Your Understanding

Problem:

An engineer builds two simple pendula. Both are suspended from small wires secured to the ceiling of a room. Each pendulum hovers 2 cm above the floor. Pendulum 1 has a bob with a mass of 10 kg. Pendulum 2 has a bob with a mass of 100 kg. Describe how the motion of the pendula will differ if the bobs are both displaced by 12^{o}.

Solution:

The movement of the pendula will not differ at all because the mass of the bob has no effect on the motion of a simple pendulum. The pendula are only affected by the period (which is related to the pendulum's length) and by the acceleration due to gravity.

Note:
PhET Explorations: Pendulum Lab
Play with one or two pendulums and discover how the period of a simple pendulum depends on the length of the string, the mass of the pendulum bob, and the amplitude of the swing. It's easy to measure the period using the photogate timer. You can vary friction and the strength of gravity. Use the pendulum to find the value of g on planet X. Notice the anharmonic behavior at large amplitude.
https://phet.colorado.edu/sims/pendulum-lab/pendulum-lab_en.html

Section Summary

- A mass m suspended by a wire of length L is a simple pendulum and undergoes simple harmonic motion for amplitudes less than about 15^{o}.

  The period of a simple pendulum is
  Equation:

$$T = 2\pi\sqrt{\frac{L}{g}},$$

where L is the length of the string and g is the acceleration due to gravity.

Conceptual Questions

Exercise:

Problem:

Pendulum clocks are made to run at the correct rate by adjusting the pendulum's length. Suppose you move from one city to another where the acceleration due to gravity is slightly greater, taking your pendulum clock with you, will you have to lengthen or shorten the pendulum to keep the correct time, other factors remaining constant? Explain your answer.

Problems & Exercises

As usual, the acceleration due to gravity in these problems is taken to be $g = 9.80 \text{ m/s}^2$, unless otherwise specified.

Exercise:

Problem:

What is the length of a pendulum that has a period of 0.500 s?

Solution:

6.21 cm

Exercise:

Problem:

Some people think a pendulum with a period of 1.00 s can be driven with “mental energy” or psycho kinetically, because its period is the same as an average heartbeat. True or not, what is the length of such a pendulum?

Exercise:

Problem: What is the period of a 1.00-m-long pendulum?

Solution:

2.01 s

Exercise:

Problem:

How long does it take a child on a swing to complete one swing if her center of gravity is 4.00 m below the pivot?

Exercise:

Problem:

The pendulum on a cuckoo clock is 5.00 cm long. What is its frequency?

Solution:

2.23 Hz

Exercise:

Problem:

Two parakeets sit on a swing with their combined center of mass 10.0 cm below the pivot. At what frequency do they swing?

Exercise:

Problem:

(a) A pendulum that has a period of 3.00000 s and that is located where the acceleration due to gravity is $9.79\ \mathrm{m/s^2}$ is moved to a location where it the acceleration due to gravity is $9.82\ \mathrm{m/s^2}$. What is its new period? (b) Explain why so many digits are needed in the value for the period, based on the relation between the period and the acceleration due to gravity.

Solution:

(a) 2.99541 s

(b) Since the period is related to the square root of the acceleration of gravity, when the acceleration changes by 1% the period changes by $(0.01)^2 = 0.01\%$ so it is necessary to have at least 4 digits after the decimal to see the changes.

Exercise:

Problem:

A pendulum with a period of 2.00000 s in one location $g = 9.80\ \mathrm{m/s^2}$ is moved to a new location where the period is now 1.99796 s. What is the acceleration due to gravity at its new location?

Exercise:

Problem:

(a) What is the effect on the period of a pendulum if you double its length?

(b) What is the effect on the period of a pendulum if you decrease its length by 5.00%?

Solution:

(a) Period increases by a factor of 1.41 ($\sqrt{2}$)

(b) Period decreases to 97.5% of old period

Exercise:

Problem:

Find the ratio of the new/old periods of a pendulum if the pendulum were transported from Earth to the Moon, where the acceleration due to gravity is 1.63 m/s^2.

Exercise:

Problem:

At what rate will a pendulum clock run on the Moon, where the acceleration due to gravity is 1.63 m/s^2, if it keeps time accurately on Earth? That is, find the time (in hours) it takes the clock's hour hand to make one revolution on the Moon.

Solution:

Slow by a factor of 2.45

Exercise:

Problem:

Suppose the length of a clock's pendulum is changed by 1.000%, exactly at noon one day. What time will it read 24.00 hours later, assuming it the pendulum has kept perfect time before the change? Note that there are two answers, and perform the calculation to four-digit precision.

Exercise:

Problem:

If a pendulum-driven clock gains 5.00 s/day, what fractional change in pendulum length must be made for it to keep perfect time?

Solution:

length must increase by 0.0116%.

Glossary

simple pendulum
: an object with a small mass suspended from a light wire or string

Energy and the Simple Harmonic Oscillator

- Determine the maximum speed of an oscillating system.

To study the energy of a simple harmonic oscillator, we first consider all the forms of energy it can have We know from Hooke’s Law: Stress and Strain Revisited that the energy stored in the deformation of a simple harmonic oscillator is a form of potential energy given by:
Equation:

$$\mathrm{PE_{el}} = \frac{1}{2}kx^2.$$

Because a simple harmonic oscillator has no dissipative forces, the other important form of energy is kinetic energy KE. Conservation of energy for these two forms is:
Equation:

$$\mathrm{KE} + \mathrm{PE_{el}} = \mathrm{constant}$$

or
Equation:

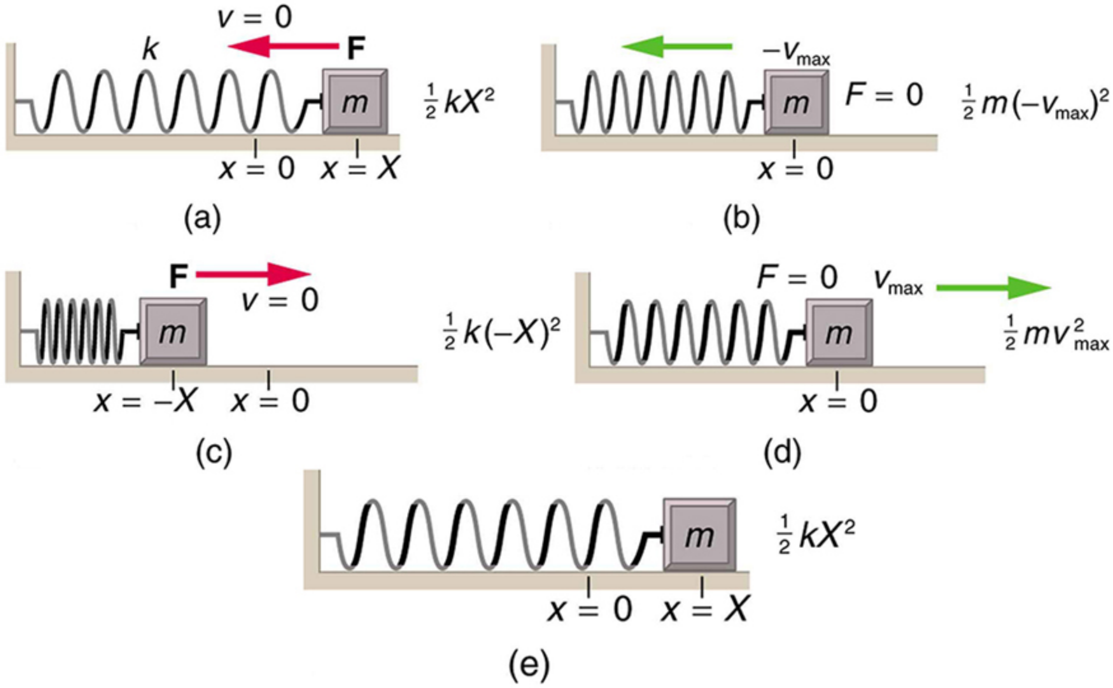
$$\frac{1}{2}\mathrm{mv}^2 + \frac{1}{2}\mathrm{kx}^2 = \mathrm{constant}.$$

This statement of conservation of energy is valid for *all* simple harmonic oscillators, including ones where the gravitational force plays a role

Namely, for a simple pendulum we replace the velocity with $v = L\omega$, the spring constant with $k = \mathrm{mg}/L$, and the displacement term with $x = L\theta$. Thus
Equation:

$$\frac{1}{2}\mathrm{mL}^2\omega^2 + \frac{1}{2}\mathrm{mgL}\theta^2 = \mathrm{constant}.$$

In the case of undamped simple harmonic motion, the energy oscillates back and forth between kinetic and potential, going completely from one to the other as the system oscillates. So for the simple example of an object on a frictionless surface attached to a spring, as shown again in [link], the motion starts with all of the energy stored in the spring. As the object starts to move, the elastic potential energy is converted to kinetic energy, becoming entirely kinetic energy at the equilibrium position. It is then converted back into elastic potential energy by the spring, the velocity becomes zero when the kinetic energy is completely converted, and so on. This concept provides extra insight here and in later applications of simple harmonic motion, such as alternating current circuits.



The transformation of energy in simple harmonic motion is illustrated for an object attached to a spring on a frictionless surface.

The conservation of energy principle can be used to derive an expression for velocity v. If we start our simple harmonic motion with zero velocity and maximum displacement ($x = X$), then the total energy is

Equation:

$$\frac{1}{2}\mathrm{kX}^2.$$

This total energy is constant and is shifted back and forth between kinetic energy and potential energy, at most times being shared by each. The conservation of energy for this system in equation form is thus:

Equation:

$$\frac{1}{2}\mathrm{mv}^2+\frac{1}{2}\mathrm{kx}^2=\frac{1}{2}\mathrm{kX}^2.$$

Solving this equation for v yields:

Equation:

$$v=\pm\sqrt{\frac{k}{m}(X^2-x^2)}.$$

Manipulating this expression algebraically gives:

Equation:

$$v=\pm\sqrt{\frac{k}{m}}X\sqrt{1-\frac{x^2}{X^2}}$$

and so

Equation:

$$v=\pm v_{\mathrm{max}}\sqrt{1-\frac{x^2}{X^2}},$$

where

Equation:

$$v_{\max} = \sqrt{\frac{k}{m}}X.$$

From this expression, we see that the velocity is a maximum ($v_{\max}$) at $x = 0$, as stated earlier in $v(t) = -v_{\max}\,\sin\frac{2\pi t}{T}$. Notice that the maximum velocity depends on three factors. Maximum velocity is directly proportional to amplitude. As you might guess, the greater the maximum displacement the greater the maximum velocity. Maximum velocity is also greater for stiffer systems, because they exert greater force for the same displacement. This observation is seen in the expression for $v_{\max}$; it is proportional to the square root of the force constant k. Finally, the maximum velocity is smaller for objects that have larger masses, because the maximum velocity is inversely proportional to the square root of m. For a given force, objects that have large masses accelerate more slowly.

A similar calculation for the simple pendulum produces a similar result, namely:

Equation:

$$\omega_{\max} = \sqrt{\frac{g}{L}}\theta_{\max}.$$

Example:

Determine the Maximum Speed of an Oscillating System: A Bumpy Road

Suppose that a car is 900 kg and has a suspension system that has a force constant $k = 6.53 \times 10^4\ \text{N/m}$. The car hits a bump and bounces with an amplitude of 0.100 m. What is its maximum vertical velocity if you assume no damping occurs?

Strategy

We can use the expression for $v_{\max}$ given in $v_{\max} = \sqrt{\frac{k}{m}}X$ to determine the maximum vertical velocity. The variables m and k are given in the

problem statement, and the maximum displacement X is 0.100 m.
Solution

1. Identify known.
2. Substitute known values into $v_{\text{max}} = \sqrt{\frac{k}{m}}X$:

   Equation:

$$v_{\text{max}} = \sqrt{\frac{6.53 \times 10^4 \text{ N/m}}{900 \text{ kg}}}(0.100 \text{ m}).$$

3. Calculate to find v_{max}= 0.852 m/s.

Discussion
This answer seems reasonable for a bouncing car. There are other ways to use conservation of energy to find v_{max}. We could use it directly, as was done in the example featured in Hooke's Law: Stress and Strain Revisited.
The small vertical displacement y of an oscillating simple pendulum, starting from its equilibrium position, is given as
Equation:

$$y(t) = a \sin \omega t,$$

where a is the amplitude, ω is the angular velocity and t is the time taken. Substituting $\omega = \frac{2\pi}{T}$, we have
Equation:

$$y(t) = a \sin\left(\frac{2\pi t}{T}\right).$$

Thus, the displacement of pendulum is a function of time as shown above.
Also the velocity of the pendulum is given by
Equation:

$$v(t) = \frac{2a\pi}{T} \cos\left(\frac{2\pi t}{T}\right),$$

so the motion of the pendulum is a function of time.

Exercise:
Check Your Understanding

Problem:

Why does it hurt more if your hand is snapped with a ruler than with a loose spring, even if the displacement of each system is equal?

Solution:

The ruler is a stiffer system, which carries greater force for the same amount of displacement. The ruler snaps your hand with greater force, which hurts more.

Exercise:
Check Your Understanding

Problem:

You are observing a simple harmonic oscillator. Identify one way you could decrease the maximum velocity of the system.

Solution:

You could increase the mass of the object that is oscillating.

Section Summary

- Energy in the simple harmonic oscillator is shared between elastic potential energy and kinetic energy, with the total being constant:
  Equation:

$$\frac{1}{2}\mathrm{mv}^2 + \frac{1}{2}\mathrm{kx}^2 = \mathrm{constant}.$$

- Maximum velocity depends on three factors: it is directly proportional to amplitude, it is greater for stiffer systems, and it is smaller for objects that have larger masses:
  Equation:

$$v_{\text{max}} = \sqrt{\frac{k}{m}}X.$$

Conceptual Questions

Exercise:

Problem:

Explain in terms of energy how dissipative forces such as friction reduce the amplitude of a harmonic oscillator. Also explain how a driving mechanism can compensate. (A pendulum clock is such a system.)

Problems & Exercises

Exercise:

Problem:

The length of nylon rope from which a mountain climber is suspended has a force constant of $1.40 \times 10^4 \text{ N/m}$.

(a) What is the frequency at which he bounces, given his mass plus and the mass of his equipment are 90.0 kg?

(b) How much would this rope stretch to break the climber’s fall if he free-falls 2.00 m before the rope runs out of slack? Hint: Use conservation of energy.

(c) Repeat both parts of this problem in the situation where twice this length of nylon rope is used.

Solution:

(a) 1.99 Hz

(b) 50.2 cm

(c) 1.41 Hz, 0.710 m

Exercise:

Problem: **Engineering Application**

Near the top of the Citigroup Center building in New York City, there is an object with mass of 4.00×10^5 kg on springs that have adjustable force constants. Its function is to dampen wind-driven oscillations of the building by oscillating at the same frequency as the building is being driven—the driving force is transferred to the object, which oscillates instead of the entire building. (a) What effective force constant should the springs have to make the object oscillate with a period of 2.00 s? (b) What energy is stored in the springs for a 2.00-m displacement from equilibrium?

Solution:

(a) $3.95 \times 10^6 \text{ N/m}$

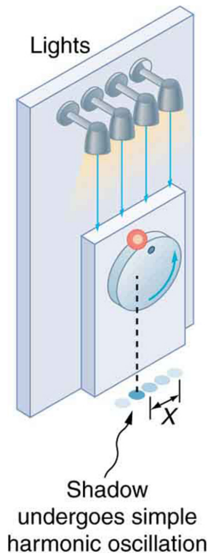
(b) $7.90 \times 10^6 \text{ J}$

Uniform Circular Motion and Simple Harmonic Motion

- Compare simple harmonic motion with uniform circular motion.

The horses on this merry-go-round exhibit uniform circular motion. (credit: Wonderlane, Flickr)

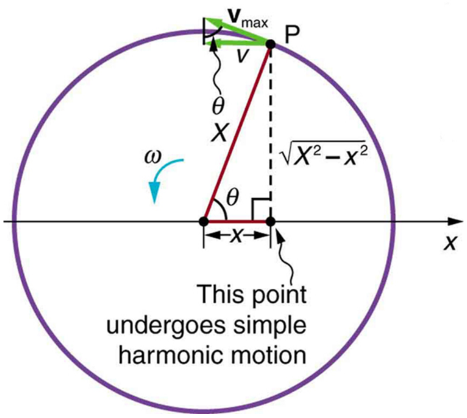
There is an easy way to produce simple harmonic motion by using uniform circular motion. [link] shows one way of using this method. A ball is attached to a uniformly rotating vertical turntable, and its shadow is projected on the floor as shown. The shadow undergoes simple harmonic motion. Hooke's law usually describes uniform circular motions (ω constant) rather than systems that have large visible displacements. So observing the projection of uniform circular motion, as in [link], is often easier than observing a precise large-scale simple harmonic oscillator. If studied in sufficient depth, simple harmonic motion produced in this manner can give considerable insight into many aspects of oscillations and waves and is very useful mathematically. In our brief treatment, we shall indicate some of the major features of this relationship and how they might be useful.



The shadow of a ball rotating at constant angular velocity ω on a turntable goes back and forth in precise simple harmonic motion.

[link] shows the basic relationship between uniform circular motion and simple harmonic motion. The point P travels around the circle at constant angular velocity ω. The point P is analogous to an object on the merry-go-

round. The projection of the position of P onto a fixed axis undergoes simple harmonic motion and is analogous to the shadow of the object. At the time shown in the figure, the projection has position x and moves to the left with velocity v. The velocity of the point P around the circle equals v .The projection of v on the x-axis is the velocity v of the simple harmonic motion along the x-axis.



A point P moving on a circular path with a constant angular velocity ω is undergoing uniform circular motion. Its projection on the x-axis undergoes simple harmonic motion. Also shown is the velocity of this point around the circle, v , and its projection, which is v. Note that these velocities form a similar triangle to the displacement triangle.

To see that the projection undergoes simple harmonic motion, note that its position x is given by

Equation:

$$x \quad X \quad \theta$$

where $\theta \quad \omega t$, ω is the constant angular velocity, and X is the radius of the circular path. Thus,

Equation:

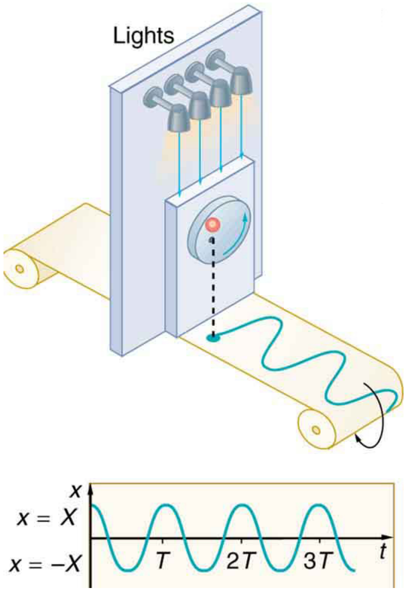
$$x \quad X \quad \omega t$$

The angular velocity ω is in radians per unit time; in this case π radians is the time for one revolution T. That is, $\omega \quad \pi \; T$. Substituting this expression for ω, we see that the position x is given by:

Equation:

$$x \; t \qquad \frac{\pi t}{T}$$

This expression is the same one we had for the position of a simple harmonic oscillator in Simple Harmonic Motion: A Special Periodic Motion. If we make a graph of position versus time as in [link], we see again the wavelike character (typical of simple harmonic motion) of the projection of uniform circular motion onto the x-axis.



The position of the projection of uniform circular motion performs simple harmonic motion, as this wavelike graph of x versus t indicates.

Now let us use [link] to do some further analysis of uniform circular motion as it relates to simple harmonic motion. The triangle formed by the velocities in the figure and the triangle formed by the displacements ($X \quad x$ and $\overline{X \quad\quad x}$) are similar right triangles. Taking ratios of similar sides, we see that

Equation:

$$\frac{v}{v} \qquad \frac{\overline{X \quad\quad x}}{X} \qquad \overline{\quad\quad\quad} \frac{x}{X}$$

We can solve this equation for the speed v or
Equation:

$$v \quad v \qquad \sqrt{\qquad \frac{x}{X}}$$

This expression for the speed of a simple harmonic oscillator is exactly the same as the equation obtained from conservation of energy considerations in [Energy and the Simple Harmonic Oscillator](#).You can begin to see that it is possible to get all of the characteristics of simple harmonic motion from an analysis of the projection of uniform circular motion.

Finally, let us consider the period T of the motion of the projection. This period is the time it takes the point P to complete one revolution. That time is the circumference of the circle πX divided by the velocity around the circle, v . Thus, the period T is
Equation:

$$T \quad \frac{\pi}{v}$$

We know from conservation of energy considerations that
Equation:

$$v \qquad \sqrt{\frac{k}{m}}X$$

Solving this equation for $X \; v$ gives
Equation:

$$\frac{X}{v} \qquad \sqrt{\frac{m}{k}}$$

Substituting this expression into the equation for T yields

Equation:

$$T \qquad \pi \ \overline{\frac{m}{k}}$$

Thus, the period of the motion is the same as for a simple harmonic oscillator. We have determined the period for any simple harmonic oscillator using the relationship between uniform circular motion and simple harmonic motion.

Some modules occasionally refer to the connection between uniform circular motion and simple harmonic motion. Moreover, if you carry your study of physics and its applications to greater depths, you will find this relationship useful. It can, for example, help to analyze how waves add when they are superimposed.

Exercise:

Check Your Understanding

Problem:

Identify an object that undergoes uniform circular motion. Describe how you could trace the simple harmonic motion of this object as a wave.

Solution:

A record player undergoes uniform circular motion. You could attach dowel rod to one point on the outside edge of the turntable and attach a pen to the other end of the dowel. As the record player turns, the pen will move. You can drag a long piece of paper under the pen, capturing its motion as a wave.

Section Summary

A projection of uniform circular motion undergoes simple harmonic oscillation.

Problems & Exercises

Exercise:

Problem:

(a)What is the maximum velocity of an 85.0-kg person bouncing on a bathroom scale having a force constant of , if the amplitude of the bounce is 0.200 cm? (b)What is the maximum energy stored in the spring?

Solution:

a). 0.266 m/s

b). 3.00 J

Exercise:

Problem:

A novelty clock has a 0.0100-kg mass object bouncing on a spring that has a force constant of 1.25 N/m. What is the maximum velocity of the object if the object bounces 3.00 cm above and below its equilibrium position? (b) How many joules of kinetic energy does the object have at its maximum velocity?

Exercise:

Problem:

At what positions is the speed of a simple harmonic oscillator half its maximum? That is, what values of x X give v v , where X is the amplitude of the motion?

Solution:

Exercise:

Problem:

A ladybug sits 12.0 cm from the center of a Beatles music album spinning at 33.33 rpm. What is the maximum velocity of its shadow on the wall behind the turntable, if illuminated parallel to the record by the parallel rays of the setting Sun?

Damped Harmonic Motion

- Compare and discuss underdamped and overdamped oscillating systems.
- Explain critically damped system.

In order to counteract dampening forces, this mom needs to keep pushing the swing. (credit: Mohd Fazlin Mohd Effendy Ooi, Flickr)

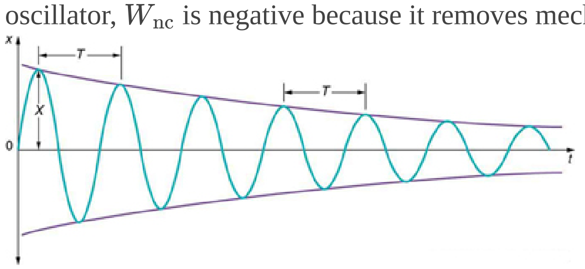
A guitar string stops oscillating a few seconds after being plucked. To keep a child happy on a swing, you must keep pushing. Although we can often make friction and other non-conservative forces negligibly small, completely undamped motion is rare. In fact, we may even want to damp oscillations, such as with car shock absorbers.

For a system that has a small amount of damping, the period and frequency are nearly the same as for simple harmonic motion, but the amplitude gradually decreases as shown in [link]. This occurs because the non-conservative damping force removes energy from the system, usually in the form of thermal energy. In general, energy removal by non-conservative forces is described as

Equation:

$$W_{\text{nc}} = \Delta(\text{KE} + \text{PE}),$$

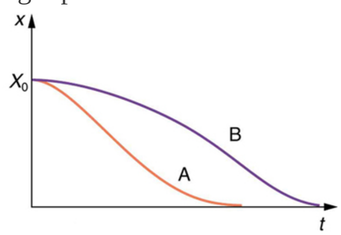
where W_{nc} is work done by a non-conservative force (here the damping force). For a damped harmonic oscillator, W_{nc} is negative because it removes mechanical energy (KE + PE) from the system.



In this graph of displacement versus time for a harmonic oscillator with a small amount of damping, the amplitude slowly decreases, but the period and frequency are nearly the

same as if the system were completely undamped.

If you gradually *increase* the amount of damping in a system, the period and frequency begin to be affected, because damping opposes and hence slows the back and forth motion. (The net force is smaller in both directions.) If there is very large damping, the system does not even oscillate—it slowly moves toward equilibrium. [link] shows the displacement of a harmonic oscillator for different amounts of damping. When we want to damp out oscillations, such as in the suspension of a car, we may want the system to return to equilibrium as quickly as possible **Critical damping** is defined as the condition in which the damping of an oscillator results in it returning as quickly as possible to its equilibrium position The critically damped system may overshoot the equilibrium position, but if it does, it will do so only once. Critical damping is represented by Curve A in [link]. With less-than critical damping, the system will return to equilibrium faster but will overshoot and cross over one or more times. Such a system is **underdamped**; its displacement is represented by the curve in [link]. Curve B in [link] represents an **overdamped** system. As with critical damping, it too may overshoot the equilibrium position, but will reach equilibrium over a longer period of time.



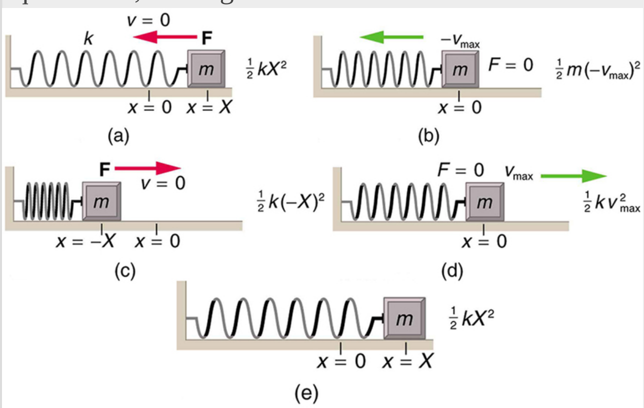
Displacement versus time for a critically damped harmonic oscillator (A) and an overdamped harmonic oscillator (B). The critically damped oscillator returns to equilibrium at $X = 0$ in the smallest time possible without overshooting.

Critical damping is often desired, because such a system returns to equilibrium rapidly and remains at equilibrium as well. In addition, a constant force applied to a critically damped system moves the system to a new equilibrium position in the shortest time possible without overshooting or oscillating about the new position. For example, when you stand on bathroom scales that have a needle gauge, the needle moves to its equilibrium position without oscillating. It would be quite inconvenient if the needle oscillated about the new equilibrium position for a long time before settling. Damping forces can vary greatly in character. Friction, for example, is sometimes independent of velocity (as assumed in most places in this text). But many damping forces depend on velocity—sometimes in complex ways, sometimes simply being proportional to velocity.

Example:

Damping an Oscillatory Motion: Friction on an Object Connected to a Spring

Damping oscillatory motion is important in many systems, and the ability to control the damping is even more so. This is generally attained using non-conservative forces such as the friction between surfaces, and viscosity for objects moving through fluids. The following example considers friction. Suppose a 0.200-kg object is connected to a spring as shown in [link], but there is simple friction between the object and the surface, and the coefficient of friction μ_k is equal to 0.0800. (a) What is the frictional force between the surfaces? (b) What total distance does the object travel if it is released 0.100 m from equilibrium, starting at $v = 0$? The force constant of the spring is $k = 50.0\ \text{N/m}$.



The transformation of energy in simple harmonic motion is illustrated for an object attached to a spring on a frictionless surface.

Strategy

This problem requires you to integrate your knowledge of various concepts regarding waves, oscillations, and damping. To solve an integrated concept problem, you must first identify the physical principles involved. Part (a) is about the frictional force. This is a topic involving the application of Newton's Laws. Part (b) requires an understanding of work and conservation of energy, as well as some understanding of horizontal oscillatory systems.

Now that we have identified the principles we must apply in order to solve the problems, we need to identify the knowns and unknowns for each part of the question, as well as the quantity that is constant in Part (a) and Part (b) of the question.

Solution a

1. Choose the proper equation: Friction is $f = \mu_\text{k}\text{mg}$.
2. Identify the known values.
3. Enter the known values into the equation:
   Equation:

$$f = (0.0800)(0.200\ \text{kg})(9.80\ \text{m/s}^2).$$

4. Calculate and convert units: $f = 0.157\ \text{N}$.

Discussion a

The force here is small because the system and the coefficients are small.

Solution b

Identify the known:

- The system involves elastic potential energy as the spring compresses and expands, friction that is related to the work done, and the kinetic energy as the body speeds up and slows down.
- Energy is not conserved as the mass oscillates because friction is a non-conservative force.
- The motion is horizontal, so gravitational potential energy does not need to be considered.
- Because the motion starts from rest, the energy in the system is initially $\mathrm{PE_{el,i}} = (1/2)\mathrm{kX^2}$. This energy is removed by work done by friction $W_{\mathrm{nc}} = -\mathrm{fd}$, where d is the total distance traveled and $f = \mu_{\mathrm{k}}\mathrm{mg}$ is the force of friction. When the system stops moving, the friction force will balance the force exerted by the spring, so $\mathrm{PE_{e1,f}} = (1/2)kx^2$ where x is the final position and is given by
  Equation:

$$\begin{aligned} F_{\mathrm{el}} &= f \\ \mathrm{kx} &= \mu_{\mathrm{k}}\mathrm{mg}. \\ x &= \frac{\mu_{\mathrm{k}}\mathrm{mg}}{k} \end{aligned}$$

1. By equating the work done to the energy removed, solve for the distance d.
2. The work done by the non-conservative forces equals the initial, stored elastic potential energy. Identify the correct equation to use:
   Equation:

$$\mathrm{W_{nc}} = \mathit{\Delta}(\mathrm{KE} + \mathrm{PE}) = \mathrm{PE_{el,f}} - \mathrm{PE_{el,i}} = \frac{1}{2}k\left(\left(\frac{\mu_{\mathrm{k}}mg}{k}\right)^2 - X^2\right).$$

3. Recall that $W_{\mathrm{nc}} = -\mathrm{fd}$.
4. Enter the friction as $f = \mu_{\mathrm{k}}\mathrm{mg}$ into $W_{\mathrm{nc}} = -\mathrm{fd}$, thus
   Equation:

$$W_{\mathrm{nc}} = -\mu_{\mathrm{k}}\mathrm{mgd}.$$

5. Combine these two equations to find
   Equation:

$$\frac{1}{2}k\left(\left(\frac{\mu_{\mathrm{k}}mg}{k}\right)^2 - X^2\right) = -\mu_{\mathrm{k}}mgd.$$

6. Solve the equation for d:
   Equation:

$$d = \frac{\mathrm{k}}{2\mu_{\mathrm{k}}\mathrm{mg}}\left(X^2 - \left(\frac{\mu_{\mathrm{k}}\mathrm{mg}}{k}\right)^2\right).$$

7. Enter the known values into the resulting equation:
   Equation:

$$d = \frac{50.0\ \mathrm{N/m}}{2(0.0800)(0.200\ \mathrm{kg})\left(9.80\ \mathrm{m/s^2}\right)}\left((0.100\ \mathrm{m})^2 - \left(\frac{(0.0800)(0.200\ \mathrm{kg})\left(9.80\ \mathrm{m/s^2}\right)}{50.0\ \mathrm{N/m}}\right)^2\right).$$

8. Calculate d and convert units:
   Equation:

$$d = 1.59\ \mathrm{m}.$$

Discussion b
This is the total distance traveled back and forth across $x = 0$, which is the undamped equilibrium position. The number of oscillations about the equilibrium position will be more than $d/X = (1.59\text{ m})/(0.100\text{ m}) = 15.9$ because the amplitude of the oscillations is decreasing with time. At the end of the motion, this system will not return to $x = 0$ for this type of damping force, because static friction will exceed the restoring force. This system is underdamped. In contrast, an overdamped system with a simple constant damping force would not cross the equilibrium position $x = 0$ a single time. For example, if this system had a damping force 20 times greater, it would only move 0.0484 m toward the equilibrium position from its original 0.100-m position.
This worked example illustrates how to apply problem-solving strategies to situations that integrate the different concepts you have learned. The first step is to identify the physical principles involved in the problem. The second step is to solve for the unknowns using familiar problem-solving strategies. These are found throughout the text, and many worked examples show how to use them for single topics. In this integrated concepts example, you can see how to apply them across several topics. You will find these techniques useful in applications of physics outside a physics course, such as in your profession, in other science disciplines, and in everyday life.

Exercise:
Check Your Understanding

Problem: Why are completely undamped harmonic oscillators so rare?

Solution:

Friction often comes into play whenever an object is moving. Friction causes damping in a harmonic oscillator.

Exercise:
Check Your Understanding

Problem: Describe the difference between overdamping, underdamping, and critical damping.

Solution:

An overdamped system moves slowly toward equilibrium. An underdamped system moves quickly to equilibrium, but will oscillate about the equilibrium point as it does so. A critically damped system moves as quickly as possible toward equilibrium without oscillating about the equilibrium.

Section Summary

- Damped harmonic oscillators have non-conservative forces that dissipate their energy.
- Critical damping returns the system to equilibrium as fast as possible without overshooting.
- An underdamped system will oscillate through the equilibrium position.
- An overdamped system moves more slowly toward equilibrium than one that is critically damped.

Conceptual Questions

Exercise:

Problem:

Give an example of a damped harmonic oscillator. (They are more common than undamped or simple harmonic oscillators.)

Exercise:

Problem: How would a car bounce after a bump under each of these conditions?

- overdamping
- underdamping
- critical damping

Exercise:

Problem:

Most harmonic oscillators are damped and, if undriven, eventually come to a stop. How is this observation related to the second law of thermodynamics?

Problems & Exercises

Exercise:

Problem:

The amplitude of a lightly damped oscillator decreases by 3.0% during each cycle. What percentage of the mechanical energy of the oscillator is lost in each cycle?

Glossary

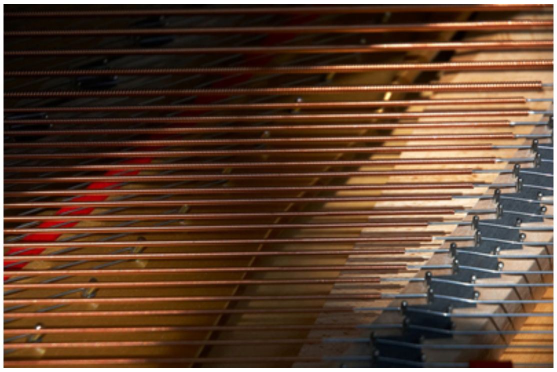
critical damping
: the condition in which the damping of an oscillator causes it to return as quickly as possible to its equilibrium position without oscillating back and forth about this position

over damping
: the condition in which damping of an oscillator causes it to return to equilibrium without oscillating; oscillator moves more slowly toward equilibrium than in the critically damped system

under damping
: the condition in which damping of an oscillator causes it to return to equilibrium with the amplitude gradually decreasing to zero; system returns to equilibrium faster but overshoots and crosses the equilibrium position one or more times

Forced Oscillations and Resonance

- Observe resonance of a paddle ball on a string.
- Observe amplitude of a damped harmonic oscillator.

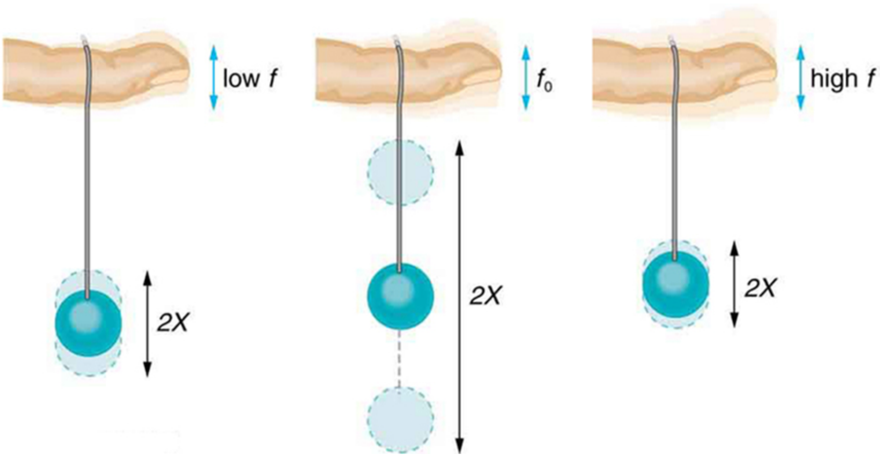


You can cause the strings in a piano to vibrate simply by producing sound waves from your voice. (credit: Matt Billings, Flickr)

Sit in front of a piano sometime and sing a loud brief note at it with the dampers off its strings. It will sing the same note back at you—the strings, having the same frequencies as your voice, are resonating in response to the forces from the sound waves that you sent to them. Your voice and a piano's strings is a good example of the fact that objects—in this case, piano strings—can be forced to oscillate but oscillate best at their natural frequency. In this section, we shall briefly explore applying a *periodic driving force* acting on a simple harmonic oscillator. The driving force puts energy into the system at a certain frequency, not necessarily the same as the natural frequency of the system. The **natural frequency** is the frequency at which a system would oscillate if there were no driving and no damping force.

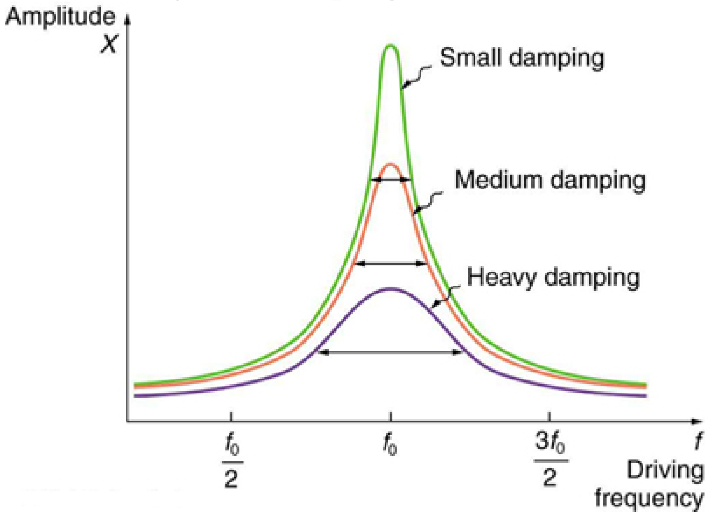
Most of us have played with toys involving an object supported on an elastic band, something like the paddle ball suspended from a finger in [link]. Imagine the finger in the figure is your finger. At first you hold your

finger steady, and the ball bounces up and down with a small amount of damping. If you move your finger up and down slowly, the ball will follow along without bouncing much on its own. As you increase the frequency at which you move your finger up and down, the ball will respond by oscillating with increasing amplitude. When you drive the ball at its natural frequency, the ball's oscillations increase in amplitude with each oscillation for as long as you drive it. The phenomenon of driving a system with a frequency equal to its natural frequency is called **resonance**. A system being driven at its natural frequency is said to **resonate**. As the driving frequency gets progressively higher than the resonant or natural frequency, the amplitude of the oscillations becomes smaller, until the oscillations nearly disappear and your finger simply moves up and down with little effect on the ball.



The paddle ball on its rubber band moves in response to the finger supporting it. If the finger moves with the natural frequency f_0 of the ball on the rubber band, then a resonance is achieved, and the amplitude of the ball's oscillations increases dramatically. At higher and lower driving frequencies, energy is transferred to the ball less efficiently, and it responds with lower-amplitude oscillations.

[link] shows a graph of the amplitude of a damped harmonic oscillator as a function of the frequency of the periodic force driving it. There are three curves on the graph, each representing a different amount of damping. All three curves peak at the point where the frequency of the driving force equals the natural frequency of the harmonic oscillator. The highest peak, or greatest response, is for the least amount of damping, because less energy is removed by the damping force.



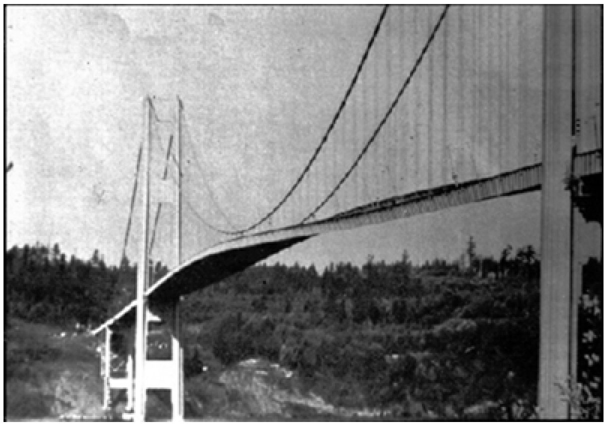
Amplitude of a harmonic oscillator as a function of the frequency of the driving force. The curves represent the same oscillator with the same natural frequency but with different amounts of damping. Resonance occurs when the driving frequency equals the natural frequency, and the greatest response is for the least amount of damping. The narrowest response is also for the least damping.

It is interesting that the widths of the resonance curves shown in [link] depend on damping: the less the damping, the narrower the resonance. The message is that if you want a driven oscillator to resonate at a very specific frequency, you need as little damping as possible. Little damping is the case

for piano strings and many other musical instruments. Conversely, if you want small-amplitude oscillations, such as in a car’s suspension system, then you want heavy damping. Heavy damping reduces the amplitude, but the tradeoff is that the system responds at more frequencies.

These features of driven harmonic oscillators apply to a huge variety of systems. When you tune a radio, for example, you are adjusting its resonant frequency so that it only oscillates to the desired station’s broadcast (driving) frequency. The more selective the radio is in discriminating between stations, the smaller its damping. Magnetic resonance imaging (MRI) is a widely used medical diagnostic tool in which atomic nuclei (mostly hydrogen nuclei) are made to resonate by incoming radio waves (on the order of 100 MHz). A child on a swing is driven by a parent at the swing’s natural frequency to achieve maximum amplitude. In all of these cases, the efficiency of energy transfer from the driving force into the oscillator is best at resonance. Speed bumps and gravel roads prove that even a car’s suspension system is not immune to resonance. In spite of finely engineered shock absorbers, which ordinarily convert mechanical energy to thermal energy almost as fast as it comes in, speed bumps still cause a large-amplitude oscillation. On gravel roads that are corrugated, you may have noticed that if you travel at the “wrong” speed, the bumps are very noticeable whereas at other speeds you may hardly feel the bumps at all. [link] shows a photograph of a famous example (the Tacoma Narrows Bridge) of the destructive effects of a driven harmonic oscillation. The Millennium Bridge in London was closed for a short period of time for the same reason while inspections were carried out.

In our bodies, the chest cavity is a clear example of a system at resonance. The diaphragm and chest wall drive the oscillations of the chest cavity which result in the lungs inflating and deflating. The system is critically damped and the muscular diaphragm oscillates at the resonant value for the system, making it highly efficient.



In 1940, the Tacoma Narrows Bridge in Washington state collapsed. Heavy cross winds drove the bridge into oscillations at its resonant frequency. Damping decreased when support cables broke loose and started to slip over the towers, allowing increasingly greater amplitudes until the structure failed (credit: PRI's *Studio 360*, via Flickr)

Exercise:
Check Your Understanding

Problem:

A famous magic trick involves a performer singing a note toward a crystal glass until the glass shatters. Explain why the trick works in terms of resonance and natural frequency.

Solution:

The performer must be singing a note that corresponds to the natural frequency of the glass. As the sound wave is directed at the glass, the glass responds by resonating at the same frequency as the sound wave.

With enough energy introduced into the system, the glass begins to vibrate and eventually shatters.

Section Summary

- A system’s natural frequency is the frequency at which the system will oscillate if not affected by driving or damping forces.
- A periodic force driving a harmonic oscillator at its natural frequency produces resonance. The system is said to resonate.
- The less damping a system has, the higher the amplitude of the forced oscillations near resonance. The more damping a system has, the broader response it has to varying driving frequencies.

Conceptual Questions

Exercise:

Problem:

Why are soldiers in general ordered to “route step” (walk out of step) across a bridge?

Problems & Exercises

Exercise:

Problem:

How much energy must the shock absorbers of a 1200-kg car dissipate in order to damp a bounce that initially has a velocity of 0.800 m/s at the equilibrium position? Assume the car returns to its original vertical position.

Solution:

384 J

Exercise:

Problem:

If a car has a suspension system with a force constant of $5.00 \times 10^4\ \mathrm{N/m}$, how much energy must the car's shocks remove to dampen an oscillation starting with a maximum displacement of 0.0750 m?

Exercise:

Problem:

(a) How much will a spring that has a force constant of 40.0 N/m be stretched by an object with a mass of 0.500 kg when hung motionless from the spring? (b) Calculate the decrease in gravitational potential energy of the 0.500-kg object when it descends this distance. (c) Part of this gravitational energy goes into the spring. Calculate the energy stored in the spring by this stretch, and compare it with the gravitational potential energy. Explain where the rest of the energy might go.

Solution:

(a). 0.123 m

(b). −0.600 J

(c). 0.300 J. The rest of the energy may go into heat caused by friction and other damping forces.

Exercise:

Problem:

Suppose you have a 0.750-kg object on a horizontal surface connected to a spring that has a force constant of 150 N/m. There is simple friction between the object and surface with a static coefficient of friction $\mu_s = 0.100$. (a) How far can the spring be stretched without moving the mass? (b) If the object is set into oscillation with an amplitude twice the distance found in part (a), and the kinetic coefficient of friction is $\mu_k = 0.0850$, what total distance does it travel before stopping? Assume it starts at the maximum amplitude.

Exercise:

Problem:

Engineering Application: A suspension bridge oscillates with an effective force constant of 1.00×10^8 N/m. (a) How much energy is needed to make it oscillate with an amplitude of 0.100 m? (b) If soldiers march across the bridge with a cadence equal to the bridge's natural frequency and impart 1.00×10^4 J of energy each second, how long does it take for the bridge's oscillations to go from 0.100 m to 0.500 m amplitude?

Solution:

(a) 5.00×10^5 J

(b) 1.20×10^3 s

Glossary

natural frequency
: the frequency at which a system would oscillate if there were no driving and no damping forces

resonance

the phenomenon of driving a system with a frequency equal to the system's natural frequency

resonate
a system being driven at its natural frequency

Waves

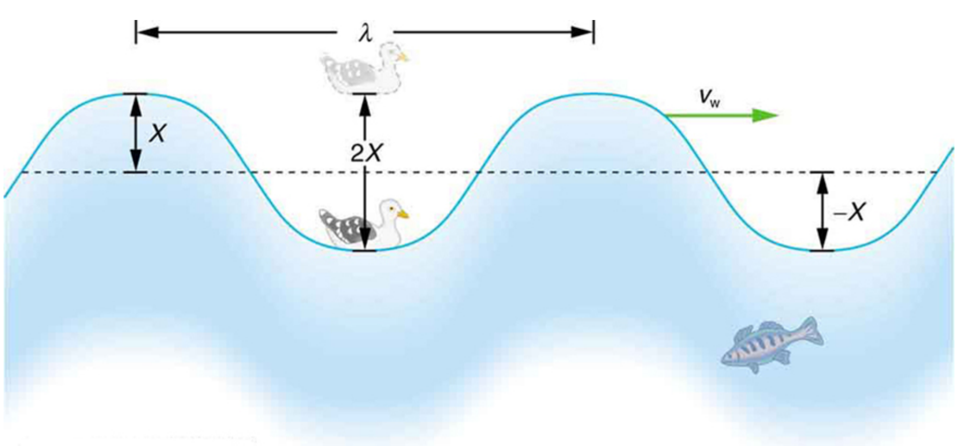
- State the characteristics of a wave.
- Calculate the velocity of wave propagation.

Waves in the ocean behave similarly to all other types of waves. (credit: Steve Jurveston, Flickr)

What do we mean when we say something is a wave? The most intuitive and easiest wave to imagine is the familiar water wave. More precisely, a **wave** is a disturbance that propagates, or moves from the place it was created. For water waves, the disturbance is in the surface of the water, perhaps created by a rock thrown into a pond or by a swimmer splashing the surface repeatedly. For sound waves, the disturbance is a change in air pressure, perhaps created by the oscillating cone inside a speaker. For earthquakes, there are several types of disturbances, including disturbance of Earth's surface and pressure disturbances under the surface. Even radio waves are most easily understood using an analogy with water waves. Visualizing water waves is useful because there is more to it than just a mental image. Water waves exhibit characteristics common to all waves, such as amplitude, period, frequency and energy. All wave characteristics can be described by a small set of underlying principles.

A wave is a disturbance that propagates, or moves from the place it was created. The simplest waves repeat themselves for several cycles and are associated with simple harmonic motion. Let us start by considering the simplified water wave in [link]. The wave is an up and down disturbance of the water surface. It causes a sea gull to move up and down in simple harmonic motion as the wave crests and troughs (peaks and valleys) pass under the bird. The time for one complete up and down motion is the wave's period T. The wave's frequency is $f=1/T$, as usual. The wave itself moves to the right in the figure. This movement of the wave is actually the disturbance moving to the right, not the water itself (or the bird would move to the right). We define **wave velocity** v_{w} to be the speed at which the disturbance moves. Wave velocity is sometimes also called the *propagation velocity or propagation speed*, because the disturbance propagates from one location to another.

Note:
Misconception Alert
Many people think that water waves push water from one direction to another. In fact, the particles of water tend to stay in one location, save for moving up and down due to the energy in the wave. The energy moves forward through the water, but the water stays in one place. If you feel yourself pushed in an ocean, what you feel is the energy of the wave, not a rush of water.



An idealized ocean wave passes under a sea gull that bobs up and down in simple harmonic motion. The wave has a wavelength λ, which is the distance between adjacent identical parts of the wave. The up and down disturbance of the surface propagates parallel to the surface at a speed v_w.

The water wave in the figure also has a length associated with it, called its **wavelength** λ, the distance between adjacent identical parts of a wave. (λ is the distance parallel to the direction of propagation.) The speed of propagation v_w is the distance the wave travels in a given time, which is one wavelength in the time of one period. In equation form, that is

Equation:

$$v_w = \frac{\lambda}{T}$$

or

Equation:

$$v_w = f\lambda.$$

This fundamental relationship holds for all types of waves. For water waves, v_w is the speed of a surface wave; for sound, v_w is the speed of sound; and for visible light, v_w is the speed of light, for example.

Note:

Take-Home Experiment: Waves in a Bowl

Fill a large bowl or basin with water and wait for the water to settle so there are no ripples. Gently drop a cork into the middle of the bowl. Estimate the wavelength and period of oscillation of the water wave that propagates away from the cork. Remove the cork from the bowl and wait

for the water to settle again. Gently drop the cork at a height that is different from the first drop. Does the wavelength depend upon how high above the water the cork is dropped?

Example:
Calculate the Velocity of Wave Propagation: Gull in the Ocean
Calculate the wave velocity of the ocean wave in [link] if the distance between wave crests is 10.0 m and the time for a sea gull to bob up and down is 5.00 s.

Strategy
We are asked to find v_{w}. The given information tells us that $\lambda = 10.0\text{ m}$ and $T = 5.00\text{ s}$. Therefore, we can use $v_{\text{w}} = \frac{\lambda}{T}$ to find the wave velocity.

Solution

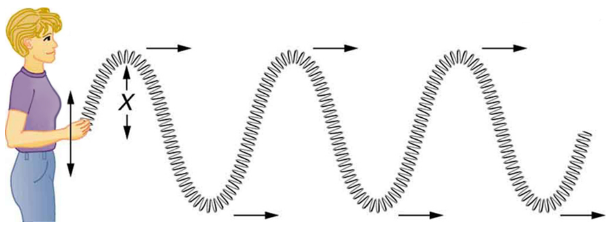
1. Enter the known values into $v_{\text{w}} = \frac{\lambda}{T}$:
   Equation:
   $$v_{\text{w}} = \frac{10.0\text{ m}}{5.00\text{ s}}.$$
2. Solve for v_{w} to find v_{w}= 2.00 m/s.

Discussion
This slow speed seems reasonable for an ocean wave. Note that the wave moves to the right in the figure at this speed, not the varying speed at which the sea gull moves up and down.

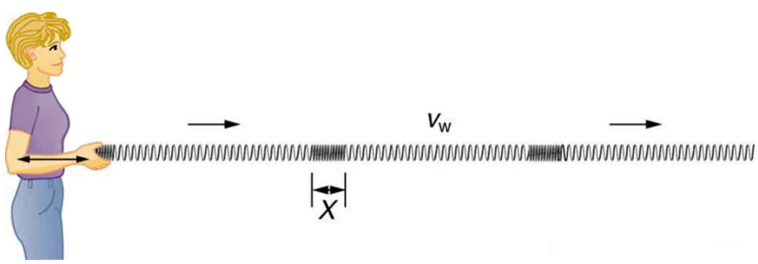
Transverse and Longitudinal Waves

A simple wave consists of a periodic disturbance that propagates from one place to another. The wave in [link] propagates in the horizontal direction while the surface is disturbed in the vertical direction. Such a wave is called a **transverse wave** or shear wave; in such a wave, the disturbance is perpendicular to the direction of propagation. In contrast, in a **longitudinal**

wave or compressional wave, the disturbance is parallel to the direction of propagation. [link] shows an example of a longitudinal wave. The size of the disturbance is its amplitude X and is completely independent of the speed of propagation v_w.



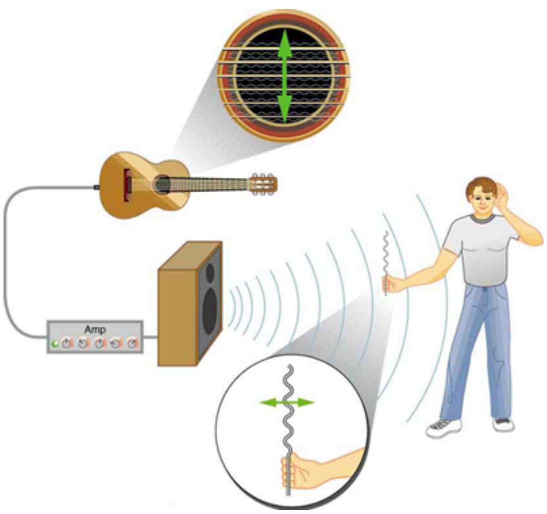
In this example of a transverse wave, the wave propagates horizontally, and the disturbance in the cord is in the vertical direction.



In this example of a longitudinal wave, the wave propagates horizontally, and the disturbance in the cord is also in the horizontal direction.

Waves may be transverse, longitudinal, or *a combination of the two*. (Water waves are actually a combination of transverse and longitudinal. The simplified water wave illustrated in [link] shows no longitudinal motion of the bird.) The waves on the strings of musical instruments are transverse—so are electromagnetic waves, such as visible light.

Sound waves in air and water are longitudinal. Their disturbances are periodic variations in pressure that are transmitted in fluids. Fluids do not have appreciable shear strength, and thus the sound waves in them must be longitudinal or compressional. Sound in solids can be both longitudinal and transverse.



The wave on a guitar string is transverse. The sound wave rattles a sheet of paper in a direction that shows the sound wave is longitudinal.

Earthquake waves under Earth's surface also have both longitudinal and transverse components (called compressional or P-waves and shear or S-waves, respectively). These components have important individual characteristics—they propagate at different speeds, for example.

Earthquakes also have surface waves that are similar to surface waves on water.

Exercise:

Check Your Understanding

Problem:

Why is it important to differentiate between longitudinal and transverse waves?

Solution:

In the different types of waves, energy can propagate in a different direction relative to the motion of the wave. This is important to understand how different types of waves affect the materials around them.

Note:

PhET Explorations: Wave on a String

Watch a string vibrate in slow motion. Wiggle the end of the string and make waves, or adjust the frequency and amplitude of an oscillator. Adjust the damping and tension. The end can be fixed, loose, or open.

https://phet.colorado.edu/sims/html/wave-on-a-string/latest/wave-on-a-string_en.html

Section Summary

- A wave is a disturbance that moves from the point of creation with a wave velocity v_{w}.
- A wave has a wavelength λ, which is the distance between adjacent identical parts of the wave.
- Wave velocity and wavelength are related to the wave's frequency and period by $v_{\text{w}} = \frac{\lambda}{T}$ or $v_{\text{w}} = f\lambda$.

- A transverse wave has a disturbance perpendicular to its direction of propagation, whereas a longitudinal wave has a disturbance parallel to its direction of propagation.

Conceptual Questions

Exercise:

Problem:

Give one example of a transverse wave and another of a longitudinal wave, being careful to note the relative directions of the disturbance and wave propagation in each.

Exercise:

Problem:

What is the difference between propagation speed and the frequency of a wave? Does one or both affect wavelength? If so, how?

Problems & Exercises

Exercise:

Problem:

Storms in the South Pacific can create waves that travel all the way to the California coast, which are 12,000 km away. How long does it take them if they travel at 15.0 m/s?

Solution:
Equation:

$$t = 9.26\text{ d}$$

Exercise:

Problem:

Waves on a swimming pool propagate at 0.750 m/s. You splash the water at one end of the pool and observe the wave go to the opposite end, reflect, and return in 30.0 s. How far away is the other end of the pool?

Exercise:

Problem:

Wind gusts create ripples on the ocean that have a wavelength of 5.00 cm and propagate at 2.00 m/s. What is their frequency?

Solution:
Equation:

$$f = 40.0 \text{ Hz}$$

Exercise:

Problem:

How many times a minute does a boat bob up and down on ocean waves that have a wavelength of 40.0 m and a propagation speed of 5.00 m/s?

Exercise:

Problem:

Scouts at a camp shake the rope bridge they have just crossed and observe the wave crests to be 8.00 m apart. If they shake it the bridge twice per second, what is the propagation speed of the waves?

Solution:
Equation:

$$v_{\text{w}} = 16.0 \text{ m/s}$$

Exercise:

Problem:

What is the wavelength of the waves you create in a swimming pool if you splash your hand at a rate of 2.00 Hz and the waves propagate at 0.800 m/s?

Exercise:

Problem:

What is the wavelength of an earthquake that shakes you with a frequency of 10.0 Hz and gets to another city 84.0 km away in 12.0 s?

Solution:
Equation:

$$\lambda = 700 \text{ m}$$

Exercise:

Problem:

Radio waves transmitted through space at $3.00 \times 10^8 \text{ m/s}$ by the Voyager spacecraft have a wavelength of 0.120 m. What is their frequency?

Exercise:

Problem:

Your ear is capable of differentiating sounds that arrive at the ear just 1.00 ms apart. What is the minimum distance between two speakers that produce sounds that arrive at noticeably different times on a day when the speed of sound is 340 m/s?

Solution:
Equation:

$$d = 34.0 \text{ cm}$$

Exercise:

Problem:

(a) Seismographs measure the arrival times of earthquakes with a precision of 0.100 s. To get the distance to the epicenter of the quake, they compare the arrival times of S- and P-waves, which travel at different speeds. [link]) If S- and P-waves travel at 4.00 and 7.20 km/s, respectively, in the region considered, how precisely can the distance to the source of the earthquake be determined? (b) Seismic waves from underground detonations of nuclear bombs can be used to locate the test site and detect violations of test bans. Discuss whether your answer to (a) implies a serious limit to such detection. (Note also that the uncertainty is greater if there is an uncertainty in the propagation speeds of the S- and P-waves.)

A seismograph as described in above problem.(credit: Oleg Alexandrov)

Glossary

longitudinal wave

a wave in which the disturbance is parallel to the direction of propagation

transverse wave
: a wave in which the disturbance is perpendicular to the direction of propagation

wave velocity
: the speed at which the disturbance moves. Also called the propagation velocity or propagation speed

wavelength
: the distance between adjacent identical parts of a wave

Superposition and Interference

- Explain standing waves.
- Describe the mathematical representation of overtones and beat frequency.

These waves result from the superposition of several waves from different sources, producing a complex pattern. (credit: waterborough, Wikimedia Commons)

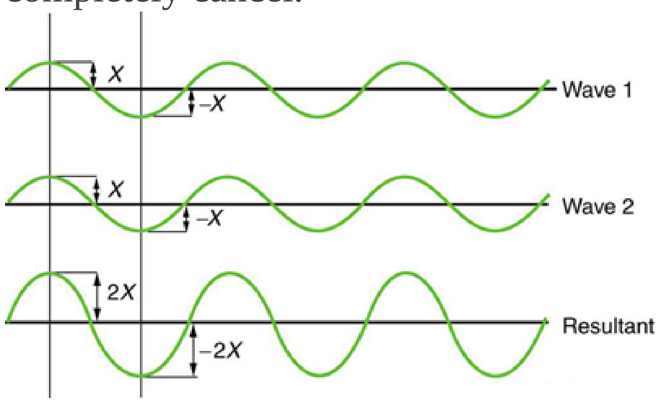
Most waves do not look very simple. They look more like the waves in [link] than like the simple water wave considered in Waves. (Simple waves may be created by a simple harmonic oscillation, and thus have a sinusoidal shape). Complex waves are more interesting, even beautiful, but they look formidable. Most waves appear complex because they result from several simple waves adding together. Luckily, the rules for adding waves are quite simple.

When two or more waves arrive at the same point, they superimpose themselves on one another. More specifically, the disturbances of waves are superimposed when they come together—a phenomenon called **superposition**. Each disturbance corresponds to a force, and forces add. If the disturbances are along the same line, then the resulting wave is a simple

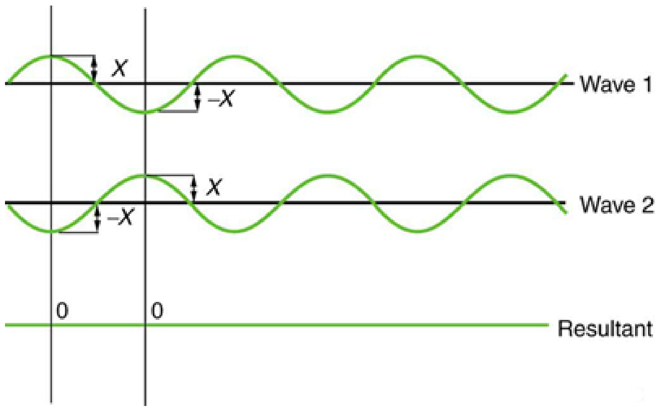
addition of the disturbances of the individual waves—that is, their amplitudes add. [link] and [link] illustrate superposition in two special cases, both of which produce simple results.

[link] shows two identical waves that arrive at the same point exactly in phase. The crests of the two waves are precisely aligned, as are the troughs. This superposition produces pure **constructive interference**. Because the disturbances add, pure constructive interference produces a wave that has twice the amplitude of the individual waves, but has the same wavelength.

[link] shows two identical waves that arrive exactly out of phase—that is, precisely aligned crest to trough—producing pure **destructive interference**. Because the disturbances are in the opposite direction for this superposition, the resulting amplitude is zero for pure destructive interference—the waves completely cancel.



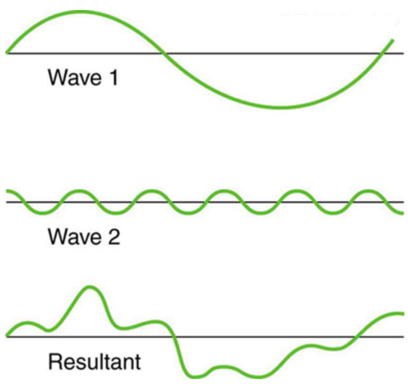
Pure constructive interference of two identical waves produces one with twice the amplitude, but the same wavelength.



Pure destructive interference of two identical waves produces zero amplitude, or complete cancellation.

While pure constructive and pure destructive interference do occur, they require precisely aligned identical waves. The superposition of most waves produces a combination of constructive and destructive interference and can vary from place to place and time to time. Sound from a stereo, for example, can be loud in one spot and quiet in another. Varying loudness means the sound waves add partially constructively and partially destructively at different locations. A stereo has at least two speakers creating sound waves, and waves can reflect from walls. All these waves superimpose. An example of sounds that vary over time from constructive to destructive is found in the combined whine of airplane jets heard by a stationary passenger. The combined sound can fluctuate up and down in volume as the sound from the two engines varies in time from constructive to destructive. These examples are of waves that are similar.

An example of the superposition of two dissimilar waves is shown in [link]. Here again, the disturbances add and subtract, producing a more complicated looking wave.

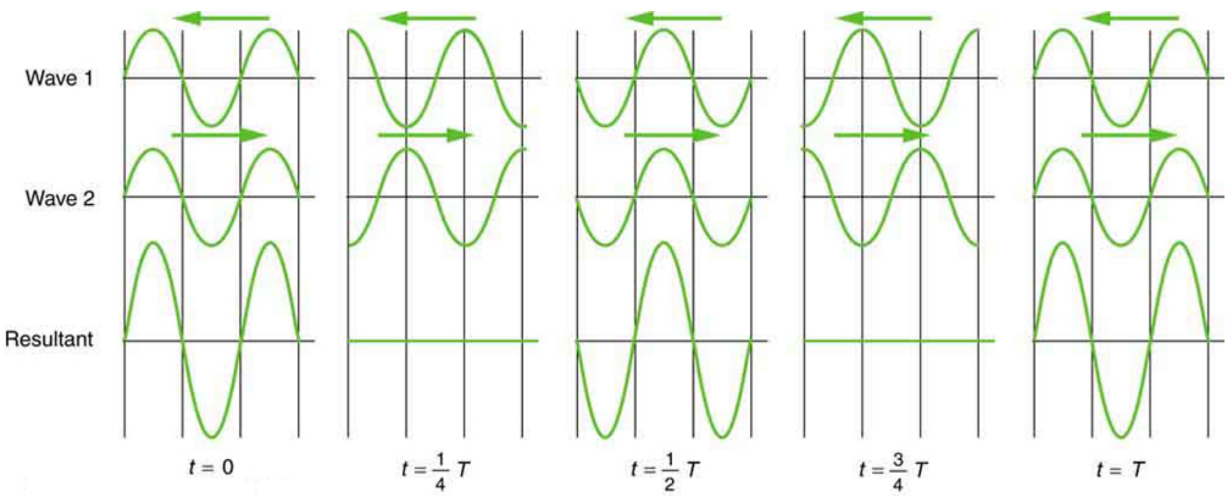


Superposition of non-identical waves exhibits both constructive and destructive interference.

Standing Waves

Sometimes waves do not seem to move; rather, they just vibrate in place. Unmoving waves can be seen on the surface of a glass of milk in a refrigerator, for example. Vibrations from the refrigerator motor create waves on the milk that oscillate up and down but do not seem to move across the surface. These waves are formed by the superposition of two or more moving waves, such as illustrated in [link] for two identical waves moving in opposite directions. The waves move through each other with their disturbances adding as they go by. If the two waves have the same amplitude and wavelength, then they alternate between constructive and destructive interference. The resultant looks like a wave standing in place and, thus, is called a **standing wave**. Waves on the glass of milk are one example of standing waves. There are other standing waves, such as on guitar strings and in organ pipes. With the glass of milk, the two waves that produce standing waves may come from reflections from the side of the glass.

A closer look at earthquakes provides evidence for conditions appropriate for resonance, standing waves, and constructive and destructive interference. A building may be vibrated for several seconds with a driving frequency matching that of the natural frequency of vibration of the building—producing a resonance resulting in one building collapsing while neighboring buildings do not. Often buildings of a certain height are devastated while other taller buildings remain intact. The building height matches the condition for setting up a standing wave for that particular height. As the earthquake waves travel along the surface of Earth and reflect off denser rocks, constructive interference occurs at certain points. Often areas closer to the epicenter are not damaged while areas farther away are damaged.

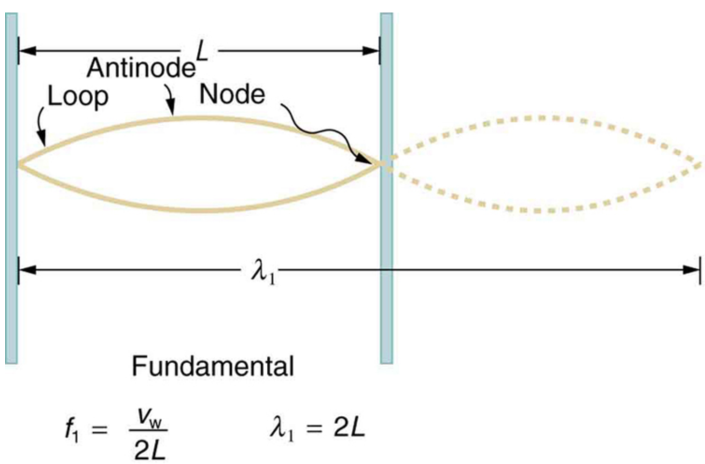


Standing wave created by the superposition of two identical waves moving in opposite directions. The oscillations are at fixed locations in space and result from alternately constructive and destructive interference.

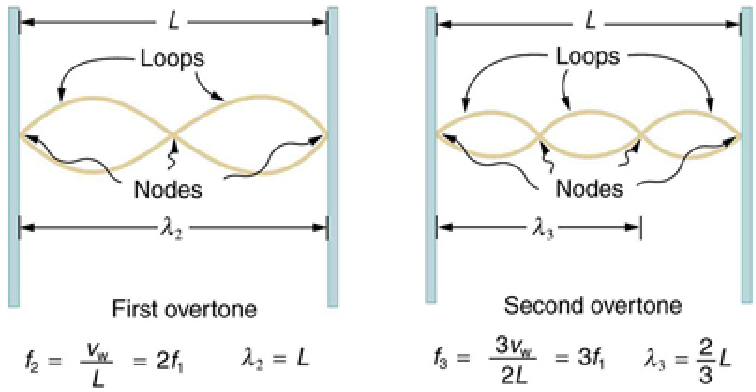
Standing waves are also found on the strings of musical instruments and are due to reflections of waves from the ends of the string. [link] and [link] show three standing waves that can be created on a string that is fixed at both ends. **Nodes** are the points where the string does not move; more

generally, nodes are where the wave disturbance is zero in a standing wave. The fixed ends of strings must be nodes, too, because the string cannot move there. The word **antinode** is used to denote the location of maximum amplitude in standing waves. Standing waves on strings have a frequency that is related to the propagation speed v_w of the disturbance on the string. The wavelength λ is determined by the distance between the points where the string is fixed in place.

The lowest frequency, called the **fundamental frequency**, is thus for the longest wavelength, which is seen to be $\lambda_1 = 2L$. Therefore, the fundamental frequency is $f_1 = v_w/\lambda_1 = v_w/2L$. In this case, the **overtones** or harmonics are multiples of the fundamental frequency. As seen in [link], the first harmonic can easily be calculated since $\lambda_2 = L$. Thus, $f_2 = v_w/\lambda_2 = v_w/2L = 2f_1$. Similarly, $f_3 = 3f_1$, and so on. All of these frequencies can be changed by adjusting the tension in the string. The greater the tension, the greater v_w is and the higher the frequencies. This observation is familiar to anyone who has ever observed a string instrument being tuned. We will see in later chapters that standing waves are crucial to many resonance phenomena, such as in sounding boxes on string instruments.



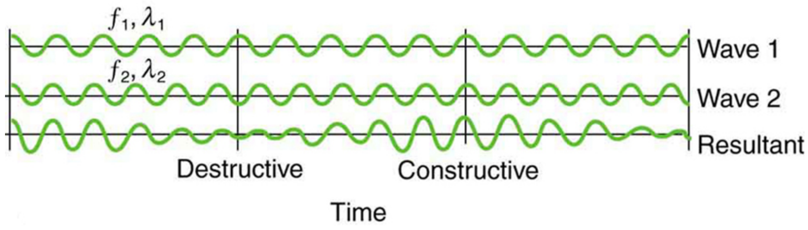
The figure shows a string oscillating at its fundamental frequency.



First and second harmonic frequencies are shown.

Beats

Striking two adjacent keys on a piano produces a warbling combination usually considered to be unpleasant. The superposition of two waves of similar but not identical frequencies is the culprit. Another example is often noticeable in jet aircraft, particularly the two-engine variety, while taxiing. The combined sound of the engines goes up and down in loudness. This varying loudness happens because the sound waves have similar but not identical frequencies. The discordant warbling of the piano and the fluctuating loudness of the jet engine noise are both due to alternately constructive and destructive interference as the two waves go in and out of phase. [link] illustrates this graphically.



Beats are produced by the superposition of two waves of slightly different frequencies but identical amplitudes. The waves alternate in time between constructive interference and destructive interference, giving the resulting wave a time-varying amplitude.

The wave resulting from the superposition of two similar-frequency waves has a frequency that is the average of the two. This wave fluctuates in amplitude, or *beats*, with a frequency called the **beat frequency**. We can determine the beat frequency by adding two waves together mathematically. Note that a wave can be represented at one point in space as

Equation:

$$x = X\cos\left(\frac{2\pi\, t}{T}\right) = X\cos(2\pi\ \mathrm{ft}),$$

where $f = 1/T$ is the frequency of the wave. Adding two waves that have different frequencies but identical amplitudes produces a resultant

Equation:

$$x = x_1 + x_2.$$

More specifically,

Equation:

$$x = X\cos(2\pi\ f_1 t) + X\cos(2\pi\ f_2 t).$$

Using a trigonometric identity, it can be shown that

Equation:

$$x = 2X\cos(\pi f_{\text{B}}t)\cos(2\pi f_{\text{ave}}t),$$

where

Equation:

$$f_{\text{B}} = \mid f_1 - f_2 \mid$$

is the beat frequency, and f_{ave} is the average of f_1 and f_2. These results mean that the resultant wave has twice the amplitude and the average frequency of the two superimposed waves, but it also fluctuates in overall amplitude at the beat frequency f_{B}. The first cosine term in the expression effectively causes the amplitude to go up and down. The second cosine term is the wave with frequency f_{ave}. This result is valid for all types of waves. However, if it is a sound wave, providing the two frequencies are similar, then what we hear is an average frequency that gets louder and softer (or warbles) at the beat frequency.

Note:

Making Career Connections

Piano tuners use beats routinely in their work. When comparing a note with a tuning fork, they listen for beats and adjust the string until the beats go away (to zero frequency). For example, if the tuning fork has a 256 Hz frequency and two beats per second are heard, then the other frequency is either 254 or 258 Hz. Most keys hit multiple strings, and these strings are actually adjusted until they have nearly the same frequency and give a slow beat for richness. Twelve-string guitars and mandolins are also tuned using beats.

While beats may sometimes be annoying in audible sounds, we will find that beats have many applications. Observing beats is a very useful way to compare similar frequencies. There are applications of beats as apparently disparate as in ultrasonic imaging and radar speed traps.

Exercise:

Check Your Understanding

Problem:

Imagine you are holding one end of a jump rope, and your friend holds the other. If your friend holds her end still, you can move your end up and down, creating a transverse wave. If your friend then begins to move her end up and down, generating a wave in the opposite direction, what resultant wave forms would you expect to see in the jump rope?

Solution:

The rope would alternate between having waves with amplitudes two times the original amplitude and reaching equilibrium with no amplitude at all. The wavelengths will result in both constructive and destructive interference

Exercise:

Check Your Understanding

Problem: Define nodes and antinodes.

Solution:

Nodes are areas of wave interference where there is no motion. Antinodes are areas of wave interference where the motion is at its maximum point.

Exercise:

Check Your Understanding

Problem:

You hook up a stereo system. When you test the system, you notice that in one corner of the room, the sounds seem dull. In another area, the sounds seem excessively loud. Describe how the sound moving about the room could result in these effects.

Solution:

With multiple speakers putting out sounds into the room, and these sounds bouncing off walls, there is bound to be some wave interference. In the dull areas, the interference is probably mostly destructive. In the louder areas, the interference is probably mostly constructive.

Note:
PhET Explorations: Wave Interference
Make waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern.

Wave Interference

Section Summary

- Superposition is the combination of two waves at the same location.
- Constructive interference occurs when two identical waves are superimposed in phase.

- Destructive interference occurs when two identical waves are superimposed exactly out of phase.
- A standing wave is one in which two waves superimpose to produce a wave that varies in amplitude but does not propagate.
- Nodes are points of no motion in standing waves.
- An antinode is the location of maximum amplitude of a standing wave.
- Waves on a string are resonant standing waves with a fundamental frequency and can occur at higher multiples of the fundamental, called overtones or harmonics.
- Beats occur when waves of similar frequencies f_1 and f_2 are superimposed. The resulting amplitude oscillates with a beat frequency given by
  Equation:

$$f_B = \mid f_1 - f_2 \mid .$$

Conceptual Questions

Exercise:

Problem:

Speakers in stereo systems have two color-coded terminals to indicate how to hook up the wires. If the wires are reversed, the speaker moves in a direction opposite that of a properly connected speaker. Explain why it is important to have both speakers connected the same way.

Problems & Exercises

Exercise:

Problem:

A car has two horns, one emitting a frequency of 199 Hz and the other emitting a frequency of 203 Hz. What beat frequency do they produce?

Solution:

$f = 4\ \text{Hz}$

Exercise:

Problem:

The middle-C hammer of a piano hits two strings, producing beats of 1.50 Hz. One of the strings is tuned to 260.00 Hz. What frequencies could the other string have?

Exercise:

Problem:

Two tuning forks having frequencies of 460 and 464 Hz are struck simultaneously. What average frequency will you hear, and what will the beat frequency be?

Solution:

462 Hz,

4 Hz

Exercise:

Problem:

Twin jet engines on an airplane are producing an average sound frequency of 4100 Hz with a beat frequency of 0.500 Hz. What are their individual frequencies?

Exercise:

Problem:

A wave traveling on a Slinky® that is stretched to 4 m takes 2.4 s to travel the length of the Slinky and back again. (a) What is the speed of the wave? (b) Using the same Slinky stretched to the same length, a standing wave is created which consists of three antinodes and four nodes. At what frequency must the Slinky be oscillating?

Solution:

(a) 3.33 m/s

(b) 1.25 Hz

Exercise:

Problem:

Three adjacent keys on a piano (F, F-sharp, and G) are struck simultaneously, producing frequencies of 349, 370, and 392 Hz. What beat frequencies are produced by this discordant combination?

Glossary

antinode
: the location of maximum amplitude in standing waves

beat frequency
: the frequency of the amplitude fluctuations of a wave

constructive interference
: when two waves arrive at the same point exactly in phase; that is, the crests of the two waves are precisely aligned, as are the troughs

destructive interference
: when two identical waves arrive at the same point exactly out of phase; that is, precisely aligned crest to trough

fundamental frequency
: the lowest frequency of a periodic waveform

nodes
: the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave

overtones

multiples of the fundamental frequency of a sound

superposition

the phenomenon that occurs when two or more waves arrive at the same point

Energy in Waves: Intensity

- Calculate the intensity and the power of rays and waves.

The destructive effect of an earthquake is palpable evidence of the energy carried in these waves. The Richter scale rating of earthquakes is related to both their amplitude and the energy they carry. (credit: Petty Officer 2nd Class Candice Villarreal, U.S. Navy)

All waves carry energy. The energy of some waves can be directly observed. Earthquakes can shake whole cities to the ground, performing the work of thousands of wrecking balls.

Loud sounds pulverize nerve cells in the inner ear, causing permanent hearing loss. Ultrasound is used for deep-heat treatment of muscle strains. A laser beam can burn away a malignancy. Water waves chew up beaches.

The amount of energy in a wave is related to its amplitude. Large-amplitude earthquakes produce large ground displacements. Loud sounds have higher pressure amplitudes and come from larger-amplitude source vibrations than

soft sounds. Large ocean breakers churn up the shore more than small ones. More quantitatively, a wave is a displacement that is resisted by a restoring force. The larger the displacement x, the larger the force $F = \mathrm{kx}$ needed to create it. Because work W is related to force multiplied by distance (Fx) and energy is put into the wave by the work done to create it, the energy in a wave is related to amplitude. In fact, a wave's energy is directly proportional to its amplitude squared because

Equation:

$$W \propto \mathrm{Fx} = \mathrm{kx}^2.$$

The energy effects of a wave depend on time as well as amplitude. For example, the longer deep-heat ultrasound is applied, the more energy it transfers. Waves can also be concentrated or spread out. Sunlight, for example, can be focused to burn wood. Earthquakes spread out, so they do less damage the farther they get from the source. In both cases, changing the area the waves cover has important effects. All these pertinent factors are included in the definition of **intensity** I as power per unit area:

Equation:

$$I = \frac{P}{A}$$

where P is the power carried by the wave through area A. The definition of intensity is valid for any energy in transit, including that carried by waves. The SI unit for intensity is watts per square meter ($\mathrm{W/m^2}$). For example, infrared and visible energy from the Sun impinge on Earth at an intensity of $1300\ \mathrm{W/m^2}$ just above the atmosphere. There are other intensity-related units in use, too. The most common is the decibel. For example, a 90 decibel sound level corresponds to an intensity of $10^{-3}\ \mathrm{W/m^2}$. (This quantity is not much power per unit area considering that 90 decibels is a relatively high sound level. Decibels will be discussed in some detail in a later chapter.

Example:
Calculating intensity and power: How much energy is in a ray of sunlight?
The average intensity of sunlight on Earth's surface is about $700\ \mathrm{W/m}^2$.
(a) Calculate the amount of energy that falls on a solar collector having an area of $0.500\ \mathrm{m}^2$ in 4.00 h.
(b) What intensity would such sunlight have if concentrated by a magnifying glass onto an area 200 times smaller than its own?
Strategy a
Because power is energy per unit time or $P = \frac{E}{t}$, the definition of intensity can be written as $I = \frac{P}{A} = \frac{E/t}{A}$, and this equation can be solved for E with the given information.
Solution a

1. Begin with the equation that states the definition of intensity:
   Equation:
   $$I = \frac{P}{A}.$$
2. Replace P with its equivalent E/t:
   Equation:
   $$I = \frac{E/t}{A}.$$
3. Solve for E:
   Equation:
   $$E = \mathrm{IAt}.$$
4. Substitute known values into the equation:
   Equation:
   $$E = \left(700\ \mathrm{W/m}^2\right)\left(0.500\ \mathrm{m}^2\right)\left[(4.00\ \mathrm{h})(3600\ \mathrm{s/h})\right].$$
5. Calculate to find E and convert units:

Equation:

$$5.04 \times 10^6 \text{ J},$$

Discussion a
The energy falling on the solar collector in 4 h in part is enough to be useful—for example, for heating a significant amount of water.
Strategy b
Taking a ratio of new intensity to old intensity and using primes for the new quantities, we will find that it depends on the ratio of the areas. All other quantities will cancel.
Solution b

1. Take the ratio of intensities, which yields:
   Equation:

$$\frac{I\prime}{I} = \frac{P\prime/A\prime}{P/A} = \frac{A}{A\prime} \left(\text{The powers cancel because } P\prime = P\right).$$

2. Identify the knowns:
   Equation:

$$A = 200A\prime,$$

   Equation:

$$\frac{I\prime}{I} = 200.$$

3. Substitute known quantities:
   Equation:

$$I\prime = 200I = 200\left(700 \text{ W/m}^2\right).$$

4. Calculate to find $I\prime$:
   Equation:

$$I\prime = 1.40 \times 10^5 \text{ W/m}^2.$$

Discussion b
Decreasing the area increases the intensity considerably. The intensity of the concentrated sunlight could even start a fire.

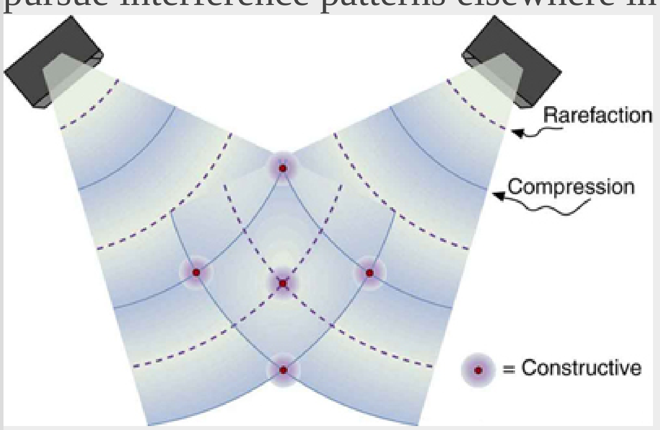
Example:
Determine the combined intensity of two waves: Perfect constructive interference
If two identical waves, each having an intensity of 1.00 W/m^2, interfere perfectly constructively, what is the intensity of the resulting wave?
Strategy
We know from Superposition and Interference that when two identical waves, which have equal amplitudes X, interfere perfectly constructively, the resulting wave has an amplitude of $2X$. Because a wave's intensity is proportional to amplitude squared, the intensity of the resulting wave is four times as great as in the individual waves.
Solution

1. Recall that intensity is proportional to amplitude squared.
2. Calculate the new amplitude:
   Equation:
   $$I\prime \propto (X\prime)^2 = (2X)^2 = 4X^2.$$
3. Recall that the intensity of the old amplitude was:
   Equation:
   $$I \propto X^2.$$
4. Take the ratio of new intensity to the old intensity. This gives:
   Equation:
   $$\frac{I\prime}{I} = 4.$$

5. Calculate to find $I\prime$:
   Equation:

$$I\prime = 4I = 4.00\ \text{W/m}^2.$$

Discussion
The intensity goes up by a factor of 4 when the amplitude doubles. This answer is a little disquieting. The two individual waves each have intensities of $1.00\ \text{W/m}^2$, yet their sum has an intensity of $4.00\ \text{W/m}^2$, which may appear to violate conservation of energy. This violation, of course, cannot happen. What does happen is intriguing. The area over which the intensity is $4.00\ \text{W/m}^2$ is much less than the area covered by the two waves before they interfered. There are other areas where the intensity is zero. The addition of waves is not as simple as our first look in Superposition and Interference suggested. We actually get a pattern of both constructive interference and destructive interference whenever two waves are added. For example, if we have two stereo speakers putting out $1.00\ \text{W/m}^2$ each, there will be places in the room where the intensity is $4.00\ \text{W/m}^2$, other places where the intensity is zero, and others in between. [link] shows what this interference might look like. We will pursue interference patterns elsewhere in this text.



These stereo speakers produce both constructive interference and destructive interference in the room, a property common to the

superposition of all types of waves. The shading is proportional to intensity.

Exercise:
Check Your Understanding

Problem:

Which measurement of a wave is most important when determining the wave's intensity?

Solution:

Amplitude, because a wave’s energy is directly proportional to its amplitude squared.

Section Summary

Intensity is defined to be the power per unit area:

$I = \frac{P}{A}$ and has units of $\mathrm{W/m^2}$.

Conceptual Questions

Exercise:

Problem:

Two identical waves undergo pure constructive interference. Is the resultant intensity twice that of the individual waves? Explain your answer.

Exercise:

Problem:

Circular water waves decrease in amplitude as they move away from where a rock is dropped. Explain why.

Problems & Exercises

Exercise:

Problem: **Medical Application**

Ultrasound of intensity $1.50 \times 10^2 \text{ W/m}^2$ is produced by the rectangular head of a medical imaging device measuring 3.00 by 5.00 cm. What is its power output?

Solution:

0.225 W

Exercise:

Problem:

The low-frequency speaker of a stereo set has a surface area of 0.05 m^2 and produces 1W of acoustical power. What is the intensity at the speaker? If the speaker projects sound uniformly in all directions, at what distance from the speaker is the intensity 0.1 W/m^2?

Exercise:

Problem:

To increase intensity of a wave by a factor of 50, by what factor should the amplitude be increased?

Solution:

7.07

Exercise:

Problem: **Engineering Application**

A device called an insolation meter is used to measure the intensity of sunlight has an area of 100 cm^2 and registers 6.50 W. What is the intensity in $\mathrm{W/m^2}$?

Exercise:

Problem: **Astronomy Application**

Energy from the Sun arrives at the top of the Earth's atmosphere with an intensity of $1.30\ \mathrm{kW/m^2}$. How long does it take for $1.8 \times 10^9\ \mathrm{J}$ to arrive on an area of $1.00\ \mathrm{m^2}$?

Solution:

16.0 d

Exercise:

Problem:

Suppose you have a device that extracts energy from ocean breakers in direct proportion to their intensity. If the device produces 10.0 kW of power on a day when the breakers are 1.20 m high, how much will it produce when they are 0.600 m high?

Solution:

2.50 kW

Exercise:

Problem: **Engineering Application**

(a) A photovoltaic array of (solar cells) is 10.0% efficient in gathering solar energy and converting it to electricity. If the average intensity of

sunlight on one day is $700\ \mathrm{W/m^2}$, what area should your array have to gather energy at the rate of 100 W? (b) What is the maximum cost of the array if it must pay for itself in two years of operation averaging 10.0 hours per day? Assume that it earns money at the rate of 9.00 ¢ per kilowatt-hour.

Exercise:

Problem:

A microphone receiving a pure sound tone feeds an oscilloscope, producing a wave on its screen. If the sound intensity is originally $2.00 \times 10^{-5}\ \mathrm{W/m^2}$, but is turned up until the amplitude increases by 30.0%, what is the new intensity?

Solution:

$3.38 \times 10^{-5}\ \mathrm{W/m^2}$

Exercise:

Problem: Medical Application

(a) What is the intensity in $\mathrm{W/m^2}$ of a laser beam used to burn away cancerous tissue that, when 90.0% absorbed, puts 500 J of energy into a circular spot 2.00 mm in diameter in 4.00 s? (b) Discuss how this intensity compares to the average intensity of sunlight (about $700\ \mathrm{W/m^2}$) and the implications that would have if the laser beam entered your eye. Note how your answer depends on the time duration of the exposure.

Glossary

intensity
: power per unit area

# Introduction to the Physics of Hearing
class="introduction"

This tree fell some time ago. When it fell, atoms in the air were disturbed. Physicists would call this disturbance sound whether someone was around to hear it or not. (credit: B.A. Bowen Photography)

If a tree falls in the forest and no one is there to hear it, does it make a sound? The answer to this old philosophical question depends on how you define sound. If sound only exists when someone is around to perceive it, then there was no sound. However, if we define sound in terms of physics; that is, a disturbance of the atoms in matter transmitted from its origin outward (in other words, a wave), then there *was* a sound, even if nobody was around to hear it.

Such a wave is the physical phenomenon we call *sound.* Its perception is hearing. Both the physical phenomenon and its perception are interesting and will be considered in this text. We shall explore both sound and hearing; they are related, but are not the same thing. We will also explore the many practical uses of sound waves, such as in medical imaging.

Sound

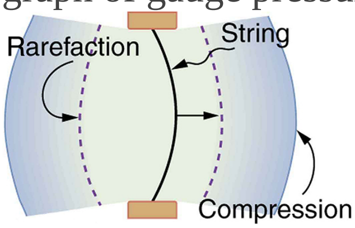
- Define sound and hearing.
- Describe sound as a longitudinal wave.

This glass has been shattered by a high-intensity sound wave of the same frequency as the resonant frequency of the glass. While the sound is not visible, the effects of the sound prove its existence. (credit: ||read||, Flickr)

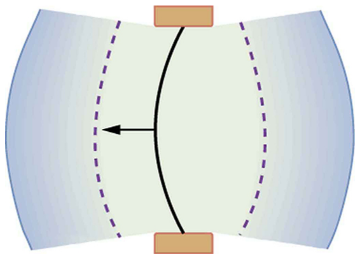
Sound can be used as a familiar illustration of waves. Because hearing is one of our most important senses, it is interesting to see how the physical properties of sound correspond to our perceptions of it. **Hearing** is the perception of sound, just as vision is the perception of visible light. But sound has important applications beyond hearing. Ultrasound, for example, is not heard but can be employed to form medical images and is also used in treatment.

The physical phenomenon of **sound** is defined to be a disturbance of matter that is transmitted from its source outward. Sound is a wave. On the atomic scale, it is a disturbance of atoms that is far more ordered than their thermal motions. In many instances, sound is a periodic wave, and the atoms undergo simple harmonic motion. In this text, we shall explore such periodic sound waves.

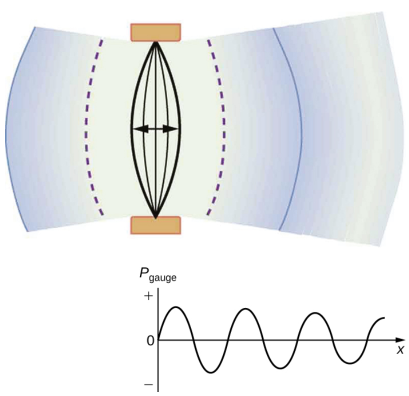
A vibrating string produces a sound wave as illustrated in [link], [link], and [link]. As the string oscillates back and forth, it transfers energy to the air, mostly as thermal energy created by turbulence. But a small part of the string's energy goes into compressing and expanding the surrounding air, creating slightly higher and lower local pressures. These compressions (high pressure regions) and rarefactions (low pressure regions) move out as longitudinal pressure waves having the same frequency as the string—they are the disturbance that is a sound wave. (Sound waves in air and most fluids are longitudinal, because fluids have almost no shear strength. In solids, sound waves can be both transverse and longitudinal.) [link] shows a graph of gauge pressure versus distance from the vibrating string.



A vibrating string moving to the right compresses the air in front of it and expands the air behind it.



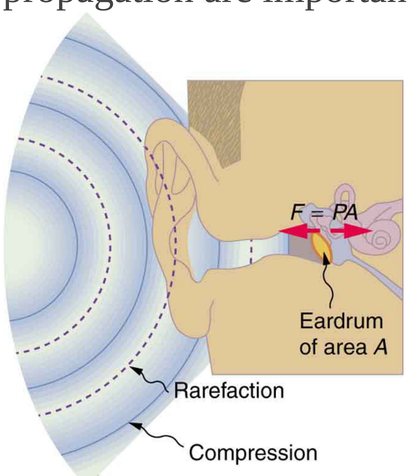
As the string moves to the left, it creates another compression and rarefaction as the ones on the right move away from the string.



After many vibrations, there are a series of compressions and rarefactions moving out from the string as a sound wave. The graph shows gauge pressure versus

distance from the source. Pressures vary only slightly from atmospheric for ordinary sounds.

The amplitude of a sound wave decreases with distance from its source, because the energy of the wave is spread over a larger and larger area. But it is also absorbed by objects, such as the eardrum in [link], and converted to thermal energy by the viscosity of air. In addition, during each compression a little heat transfers to the air and during each rarefaction even less heat transfers from the air, so that the heat transfer reduces the organized disturbance into random thermal motions. (These processes can be viewed as a manifestation of the second law of thermodynamics presented in Introduction to the Second Law of Thermodynamics: Heat Engines and Their Efficiency.) Whether the heat transfer from compression to rarefaction is significant depends on how far apart they are—that is, it depends on wavelength. Wavelength, frequency, amplitude, and speed of propagation are important for sound, as they are for all waves.



Sound wave compressions and rarefactions travel up the ear canal and

force the eardrum to vibrate. There is a net force on the eardrum, since the sound wave pressures differ from the atmospheric pressure found behind the eardrum. A complicated mechanism converts the vibrations to nerve impulses, which are perceived by the person.

Note:
PhET Explorations: Wave Interference
WMake waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern.
https://archive.cnx.org/specials/2fe7ad15-b00e-4402-b068-ff503985a18f/wave-interference/

Section Summary

- Sound is a disturbance of matter that is transmitted from its source outward.
- Sound is one type of wave.

- Hearing is the perception of sound.

Glossary

sound
: a disturbance of matter that is transmitted from its source outward

hearing
: the perception of sound

Speed of Sound, Frequency, and Wavelength

- Define pitch.
- Describe the relationship between the speed of sound, its frequency, and its wavelength.
- Describe the effects on the speed of sound as it travels through various media.
- Describe the effects of temperature on the speed of sound.

When a firework explodes, the light energy is perceived before the sound energy. Sound travels more slowly than light does. (credit: Dominic Alves, Flickr)

Sound, like all waves, travels at a certain speed and has the properties of frequency and wavelength. You can observe direct evidence of the speed of sound while watching a fireworks display. The flash of an explosion is seen well before its sound is heard, implying both that sound travels at a finite speed and that it is much slower than light. You can also directly sense the frequency of a sound. Perception of frequency is called **pitch**. The wavelength of sound is not directly sensed, but indirect evidence is found in the correlation of the size of musical instruments with their pitch. Small

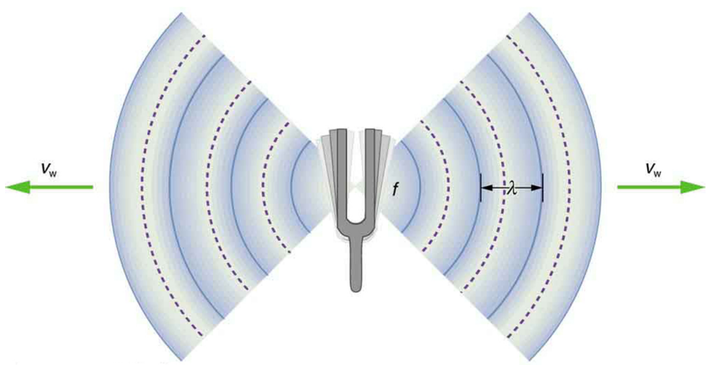
instruments, such as a piccolo, typically make high-pitch sounds, while large instruments, such as a tuba, typically make low-pitch sounds. High pitch means small wavelength, and the size of a musical instrument is directly related to the wavelengths of sound it produces. So a small instrument creates short-wavelength sounds. Similar arguments hold that a large instrument creates long-wavelength sounds.

The relationship of the speed of sound, its frequency, and wavelength is the same as for all waves:

Equation:

$$v_{\mathrm{w}} = f\lambda,$$

where v_{w} is the speed of sound, f is its frequency, and λ is its wavelength. The wavelength of a sound is the distance between adjacent identical parts of a wave—for example, between adjacent compressions as illustrated in [link]. The frequency is the same as that of the source and is the number of waves that pass a point per unit time.



A sound wave emanates from a source vibrating at a frequency f, propagates at v_{w}, and has a wavelength λ.

[link] makes it apparent that the speed of sound varies greatly in different media. The speed of sound in a medium is determined by a combination of the medium's rigidity (or compressibility in gases) and its density. The

more rigid (or less compressible) the medium, the faster the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is directly proportional to the stiffness of the oscillating object. The greater the density of a medium, the slower the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is inversely proportional to the mass of the oscillating object. The speed of sound in air is low, because air is compressible. Because liquids and solids are relatively rigid and very difficult to compress, the speed of sound in such media is generally greater than in gases.

Medium	v_w**(m/s)**
Gases at $0^{\circ}C$	
Air	331
Carbon dioxide	259
Oxygen	316
Helium	965
Hydrogen	1290
Liquids at $20^{\circ}C$	
Ethanol	1160
Mercury	1450
Water, fresh	1480

Medium	**v_w(m/s)**
Sea water	1540
Human tissue	1540
Solids (longitudinal or bulk)	
Vulcanized rubber	54
Polyethylene	920
Marble	3810
Glass, Pyrex	5640
Lead	1960
Aluminum	5120
Steel	5960

Speed of Sound in Various Media

Earthquakes, essentially sound waves in Earth's crust, are an interesting example of how the speed of sound depends on the rigidity of the medium. Earthquakes have both longitudinal and transverse components, and these travel at different speeds. The bulk modulus of granite is greater than its shear modulus. For that reason, the speed of longitudinal or pressure waves (P-waves) in earthquakes in granite is significantly higher than the speed of transverse or shear waves (S-waves). Both components of earthquakes travel slower in less rigid material, such as sediments. P-waves have speeds of 4 to 7 km/s, and S-waves correspondingly range in speed from 2 to 5 km/s, both being faster in more rigid material. The P-wave gets progressively farther ahead of the S-wave as they travel through Earth's crust. The time between the P- and S-waves is routinely used to determine the distance to their source, the epicenter of the earthquake.

The speed of sound is affected by temperature in a given medium. For air at sea level, the speed of sound is given by

Equation:

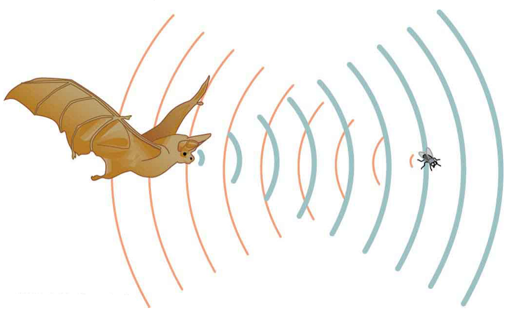
$$v_{\mathrm{w}} = (331\ \mathrm{m/s})\sqrt{\frac{T}{273\ \mathrm{K}}},$$

where the temperature (denoted as T) is in units of kelvin. The speed of sound in gases is related to the average speed of particles in the gas, v_{rms}, and that

Equation:

$$v_{\mathrm{rms}} = \sqrt{\frac{3\,kT}{m}},$$

where k is the Boltzmann constant ($1.38 \times 10^{-23}\ \mathrm{J/K}$) and m is the mass of each (identical) particle in the gas. So, it is reasonable that the speed of sound in air and other gases should depend on the square root of temperature. While not negligible, this is not a strong dependence. At $0^{\circ}\mathrm{C}$, the speed of sound is 331 m/s, whereas at $20.0^{\circ}\mathrm{C}$ it is 343 m/s, less than a 4% increase. [link] shows a use of the speed of sound by a bat to sense distances. Echoes are also used in medical imaging.



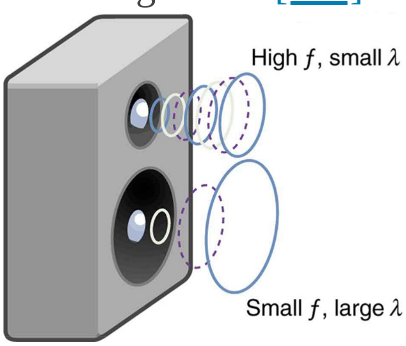
A bat uses sound echoes to find its way about and to catch prey. The time for the echo to return is directly proportional to the distance.

One of the more important properties of sound is that its speed is nearly independent of frequency. This independence is certainly true in open air for sounds in the audible range of 20 to 20,000 Hz. If this independence were not true, you would certainly notice it for music played by a marching band in a football stadium, for example. Suppose that high-frequency sounds traveled faster—then the farther you were from the band, the more the sound from the low-pitch instruments would lag that from the high-pitch ones. But the music from all instruments arrives in cadence independent of distance, and so all frequencies must travel at nearly the same speed. Recall that

Equation:

$$v_{\mathrm{w}} = f\lambda.$$

In a given medium under fixed conditions, v_{w} is constant, so that there is a relationship between f and λ; the higher the frequency, the smaller the wavelength. See [link] and consider the following example.



Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds.

Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, while the higher-frequency sounds are emitted by the small speaker, called a tweeter.

Example:
Calculating Wavelengths: What Are the Wavelengths of Audible Sounds?
Calculate the wavelengths of sounds at the extremes of the audible range, 20 and 20,000 Hz, in 30.0°C air. (Assume that the frequency values are accurate to two significant figures.)
Strategy
To find wavelength from frequency, we can use $v_{\text{w}} = f\lambda$.
Solution

1. Identify knowns. The value for v_{w}, is given by
   Equation:

$$v_{\text{w}} = (331\text{ m/s})\sqrt{\frac{T}{273\text{ K}}}.$$

2. Convert the temperature into kelvin and then enter the temperature into the equation
   Equation:

$$v_{\text{w}} = (331\text{ m/s})\sqrt{\frac{303\text{ K}}{273\text{ K}}} = 348.7\text{ m/s}.$$

3. Solve the relationship between speed and wavelength for λ:
   Equation:

$$\lambda = \frac{v_{\text{w}}}{f}.$$

4. Enter the speed and the minimum frequency to give the maximum wavelength:
   Equation:

$$\lambda_{\text{max}} = \frac{348.7 \text{ m/s}}{20 \text{ Hz}} = 17 \text{ m}.$$

5. Enter the speed and the maximum frequency to give the minimum wavelength:
   Equation:

$$\lambda_{\text{min}} = \frac{348.7 \text{ m/s}}{20{,}000 \text{ Hz}} = 0.017 \text{ m} = 1.7 \text{ cm}.$$

Discussion
Because the product of f multiplied by λ equals a constant, the smaller f is, the larger λ must be, and vice versa.

The speed of sound can change when sound travels from one medium to another. However, the frequency usually remains the same because it is like a driven oscillation and has the frequency of the original source. If v_{w} changes and f remains the same, then the wavelength λ must change. That is, because $v_{\text{w}} = f\lambda$, the higher the speed of a sound, the greater its wavelength for a given frequency.

Note:
Making Connections: Take-Home Investigation—Voice as a Sound Wave

Suspend a sheet of paper so that the top edge of the paper is fixed and the bottom edge is free to move. You could tape the top edge of the paper to the edge of a table. Gently blow near the edge of the bottom of the sheet and note how the sheet moves. Speak softly and then louder such that the sounds hit the edge of the bottom of the paper, and note how the sheet moves. Explain the effects.

Exercise:
Check Your Understanding

Problem:

Imagine you observe two fireworks explode. You hear the explosion of one as soon as you see it. However, you see the other firework for several milliseconds before you hear the explosion. Explain why this is so.

Solution:

Sound and light both travel at definite speeds. The speed of sound is slower than the speed of light. The first firework is probably very close by, so the speed difference is not noticeable. The second firework is farther away, so the light arrives at your eyes noticeably sooner than the sound wave arrives at your ears.

Exercise:
Check Your Understanding

Problem:

You observe two musical instruments that you cannot identify. One plays high-pitch sounds and the other plays low-pitch sounds. How could you determine which is which without hearing either of them play?

Solution:

Compare their sizes. High-pitch instruments are generally smaller than low-pitch instruments because they generate a smaller wavelength.

Section Summary

The relationship of the speed of sound v_{w}, its frequency f, and its wavelength λ is given by

Equation:

$$v_{\text{w}} = f\lambda,$$

which is the same relationship given for all waves.

In air, the speed of sound is related to air temperature T by

Equation:

$$v_{\text{w}} = (331\ \text{m/s})\sqrt{\frac{T}{273\ \text{K}}}.$$

v_{w} is the same for all frequencies and wavelengths.

Conceptual Questions

Exercise:

Problem:

How do sound vibrations of atoms differ from thermal motion?

Exercise:

Problem:

When sound passes from one medium to another where its propagation speed is different, does its frequency or wavelength change? Explain your answer briefly.

Problems & Exercises

Exercise:

Problem:

When poked by a spear, an operatic soprano lets out a 1200-Hz shriek. What is its wavelength if the speed of sound is 345 m/s?

Solution:

0.288 m

Exercise:

Problem:

What frequency sound has a 0.10-m wavelength when the speed of sound is 340 m/s?

Exercise:

Problem:

Calculate the speed of sound on a day when a 1500 Hz frequency has a wavelength of 0.221 m.

Solution:

332 m/s

Exercise:

Problem:

(a) What is the speed of sound in a medium where a 100-kHz frequency produces a 5.96-cm wavelength? (b) Which substance in [link] is this likely to be?

Exercise:

Problem:

Show that the speed of sound in $20.0^{\circ}C$ air is 343 m/s, as claimed in the text.

Solution:
Equation:

$$\begin{aligned} v_{\text{w}} &= (331\text{ m/s})\sqrt{\frac{T}{273\text{ K}}} = (331\text{ m/s})\sqrt{\frac{293\text{ K}}{273\text{ K}}} \\ &= 343\text{ m/s} \end{aligned}$$

Exercise:

Problem:

Air temperature in the Sahara Desert can reach $56.0^{\circ}C$ (about $134^{\circ}F$). What is the speed of sound in air at that temperature?

Exercise:

Problem:

Dolphins make sounds in air and water. What is the ratio of the wavelength of a sound in air to its wavelength in seawater? Assume air temperature is $20.0^{\circ}C$.

Solution:

0.223

Exercise:

Problem:

A sonar echo returns to a submarine 1.20 s after being emitted. What is the distance to the object creating the echo? (Assume that the submarine is in the ocean, not in fresh water.)

Exercise:

Problem:

(a) If a submarine’s sonar can measure echo times with a precision of 0.0100 s, what is the smallest difference in distances it can detect? (Assume that the submarine is in the ocean, not in fresh water.)

(b) Discuss the limits this time resolution imposes on the ability of the sonar system to detect the size and shape of the object creating the echo.

Solution:

(a) 7.70 m

(b) This means that sonar is good for spotting and locating large objects, but it isn’t able to resolve smaller objects, or detect the detailed shapes of objects. Objects like ships or large pieces of airplanes can be found by sonar, while smaller pieces must be found by other means.

Exercise:

Problem:

A physicist at a fireworks display times the lag between seeing an explosion and hearing its sound, and finds it to be 0.400 s. (a) How far away is the explosion if air temperature is 24.0°C and if you neglect the time taken for light to reach the physicist? (b) Calculate the distance to the explosion taking the speed of light into account. Note that this distance is negligibly greater.

Exercise:

Problem:

Suppose a bat uses sound echoes to locate its insect prey, 3.00 m away. (See [link].) (a) Calculate the echo times for temperatures of 5.00°C and 35.0°C. (b) What percent uncertainty does this cause for the bat in locating the insect? (c) Discuss the significance of this uncertainty and whether it could cause difficulties for the bat. (In practice, the bat continues to use sound as it closes in, eliminating most of any difficulties imposed by this and other effects, such as motion of the prey.)

Solution:

(a) 18.0 ms, 17.1 ms

(b) 5.00%

(c) This uncertainty could definitely cause difficulties for the bat, if it didn’t continue to use sound as it closed in on its prey. A 5% uncertainty could be the difference between catching the prey around the neck or around the chest, which means that it could miss grabbing its prey.

Glossary

pitch
: the perception of the frequency of a sound

Sound Intensity and Sound Level

- Define intensity, sound intensity, and sound pressure level.
- Calculate sound intensity levels in decibels (dB).

Noise on crowded roadways like this one in Delhi makes it hard to hear others unless they shout. (credit: Lingaraj G J, Flickr)

In a quiet forest, you can sometimes hear a single leaf fall to the ground. After settling into bed, you may hear your blood pulsing through your ears. But when a passing motorist has his stereo turned up, you cannot even hear what the person next to you in your car is saying. We are all very familiar with the loudness of sounds and aware that they are related to how energetically the source is vibrating. In cartoons depicting a screaming person (or an animal making a loud noise), the cartoonist often shows an open mouth with a vibrating uvula, the hanging tissue at the back of the mouth, to suggest a loud sound coming from the throat [link]. High noise exposure is hazardous to hearing, and it is common for musicians to have hearing losses that are sufficiently severe that they interfere with the musicians' abilities to perform. The relevant physical quantity is sound intensity, a concept that is valid for all sounds whether or not they are in the audible range.

Intensity is defined to be the power per unit area carried by a wave. Power is the rate at which energy is transferred by the wave. In equation form, **intensity** I is

Equation:

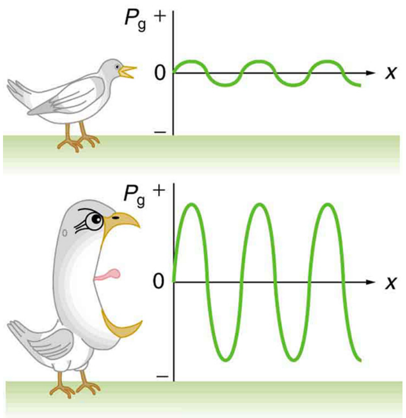
$$I = \frac{P}{A},$$

where P is the power through an area A. The SI unit for I is $\mathrm{W/m^2}$. The intensity of a sound wave is related to its amplitude squared by the following relationship:

Equation:

$$I = \frac{(\Delta p)^2}{2\rho v_{\mathrm{w}}}.$$

Here Δp is the pressure variation or pressure amplitude (half the difference between the maximum and minimum pressure in the sound wave) in units of pascals (Pa) or $\mathrm{N/m^2}$. (We are using a lower case p for pressure to distinguish it from power, denoted by P above.) The energy (as kinetic energy $\frac{mv^2}{2}$) of an oscillating element of air due to a traveling sound wave is proportional to its amplitude squared. In this equation, ρ is the density of the material in which the sound wave travels, in units of $\mathrm{kg/m^3}$, and v_{w} is the speed of sound in the medium, in units of m/s. The pressure variation is proportional to the amplitude of the oscillation, and so I varies as $(\Delta p)^2$ ([link]). This relationship is consistent with the fact that the sound wave is produced by some vibration; the greater its pressure amplitude, the more the air is compressed in the sound it creates.



Graphs of the gauge pressures in two sound waves of different intensities. The more intense sound is produced by a source that has larger-amplitude oscillations and has greater pressure maxima and minima. Because pressures are higher in the greater-intensity sound, it can exert larger forces on the objects it encounters.

Sound intensity levels are quoted in decibels (dB) much more often than sound intensities in watts per meter squared. Decibels are the unit of choice in the scientific literature as well as in the popular media. The reasons for this choice of units are related to how we perceive sounds. How our ears perceive sound can be more accurately described by the logarithm of the

intensity rather than directly to the intensity. The **sound intensity level** β in decibels of a sound having an intensity I in watts per meter squared is defined to be

Equation:

$$\beta\,(\text{dB}) = 10\,\log_{10}\left(\frac{I}{I_0}\right),$$

where $I_0 = 10^{-12}\,\text{W}/\text{m}^2$ is a reference intensity. In particular, I_0 is the lowest or threshold intensity of sound a person with normal hearing can perceive at a frequency of 1000 Hz. Sound intensity level is not the same as intensity. Because β is defined in terms of a ratio, it is a unitless quantity telling you the *level* of the sound relative to a fixed standard ($10^{-12}\,\text{W}/\text{m}^2$, in this case). The units of decibels (dB) are used to indicate this ratio is multiplied by 10 in its definition. The bel, upon which the decibel is based, is named for Alexander Graham Bell, the inventor of the telephone.

Sound intensity level β (dB)	Intensity $I(\text{W/m}^2)$	Example/effect
0	1×10^{-12}	Threshold of hearing at 1000 Hz
10	1×10^{-11}	Rustle of leaves
20	1×10^{-10}	Whisper at 1 m distance
30	1×10^{-9}	Quiet home

Sound intensity level β (dB)	Intensity $I(\mathrm{W/m^2})$	Example/effect
40	1×10^{-8}	Average home
50	1×10^{-7}	Average office, soft music
60	1×10^{-6}	Normal conversation
70	1×10^{-5}	Noisy office, busy traffic
80	1×10^{-4}	Loud radio, classroom lecture
90	1×10^{-3}	Inside a heavy truck; damage from prolonged exposure[footnote] Several government agencies and health-related professional associations recommend that 85 dB not be exceeded for 8-hour daily exposures in the absence of hearing protection.
100	1×10^{-2}	Noisy factory, siren at 30 m; damage from 8 h per day exposure
110	1×10^{-1}	Damage from 30 min per day exposure
120	1	Loud rock concert, pneumatic chipper at 2 m; threshold of pain
140	1×10^{2}	Jet airplane at 30 m; severe pain, damage in seconds
160	1×10^{4}	Bursting of eardrums

Sound Intensity Levels and Intensities

Sound Intensity Levels and Intensities

The decibel level of a sound having the threshold intensity of $10^{-12}\ \text{W/m}^2$ is $\beta = 0\ \text{dB}$, because $\log_{10}1 = 0$. That is, the threshold of hearing is 0 decibels. [link] gives levels in decibels and intensities in watts per meter squared for some familiar sounds.

One of the more striking things about the intensities in [link] is that the intensity in watts per meter squared is quite small for most sounds. The ear is sensitive to as little as a trillionth of a watt per meter squared—even more impressive when you realize that the area of the eardrum is only about $1\ \text{cm}^2$, so that only 10^{-16} W falls on it at the threshold of hearing! Air molecules in a sound wave of this intensity vibrate over a distance of less than one molecular diameter, and the gauge pressures involved are less than 10^{-9} atm.

Another impressive feature of the sounds in [link] is their numerical range. Sound intensity varies by a factor of 10^{12} from threshold to a sound that causes damage in seconds. You are unaware of this tremendous range in sound intensity because how your ears respond can be described approximately as the logarithm of intensity. Thus, sound intensity levels in decibels fit your experience better than intensities in watts per meter squared. The decibel scale is also easier to relate to because most people are more accustomed to dealing with numbers such as 0, 53, or 120 than numbers such as 1.00×10^{-11}.

One more observation readily verified by examining [link] or using $I = \frac{(\Delta p)^2}{2\rho v_{\text{w}}}$ is that each factor of 10 in intensity corresponds to 10 dB. For example, a 90 dB sound compared with a 60 dB sound is 30 dB greater, or three factors of 10 (that is, 10^3 times) as intense. Another example is that if one sound is 10^7 as intense as another, it is 70 dB higher. See [link].

I_2/I_1	$\beta_2 - \beta_1$
2.0	3.0 dB
5.0	7.0 dB
10.0	10.0 dB

Ratios of Intensities and Corresponding Differences in Sound Intensity Levels

Example:
Calculating Sound Intensity Levels: Sound Waves
Calculate the sound intensity level in decibels for a sound wave traveling in air at 0ºC and having a pressure amplitude of 0.656 Pa.
Strategy
We are given Δp, so we can calculate I using the equation $I = (\Delta p)^2 / (2pv_{\mathrm{w}})^2$. Using I, we can calculate β straight from its definition in $\beta\ (\mathrm{dB}) = 10\ \log_{10}(I/I_0)$.
Solution
(1) Identify knowns:
Sound travels at 331 m/s in air at 0ºC.
Air has a density of $1.29\ \mathrm{kg/m^3}$ at atmospheric pressure and 0ºC.
(2) Enter these values and the pressure amplitude into $I = (\Delta p)^2 /\ (2\rho v_{\mathrm{w}})$:
Equation:

$$I = \frac{(\Delta p)^2}{2\rho v_{\mathrm{w}}} = \frac{(0.656\ \mathrm{Pa})^2}{2\left(1.29\ \mathrm{kg/m^3}\right)(331\ \mathrm{m/s})} = 5.04 \times 10^{-4}\ \mathrm{W/m^2}.$$

(3) Enter the value for I and the known value for I_0 into $\beta\ (\mathrm{dB}) = 10\ \log_{10}(I/I_0)$. Calculate to find the sound intensity level in decibels:
Equation:

$$10\ \log_{10}\left(5.04 \times 10^{8}\right) = 10\ \left(8.70\right)\ \text{dB} = 87\ \text{dB}.$$

Discussion
This 87 dB sound has an intensity five times as great as an 80 dB sound. So a factor of five in intensity corresponds to a difference of 7 dB in sound intensity level. This value is true for any intensities differing by a factor of five.

Example:
Change Intensity Levels of a Sound: What Happens to the Decibel Level?
Show that if one sound is twice as intense as another, it has a sound level about 3 dB higher.
Strategy
You are given that the ratio of two intensities is 2 to 1, and are then asked to find the difference in their sound levels in decibels. You can solve this problem using of the properties of logarithms.
Solution
(1) Identify knowns:
The ratio of the two intensities is 2 to 1, or:
Equation:

$$\frac{I_2}{I_1} = 2.00.$$

We wish to show that the difference in sound levels is about 3 dB. That is, we want to show:
Equation:

$$\beta_2 - \beta_1 = 3\ \text{dB}.$$

Note that:
Equation:

$$\log_{10} b - \log_{10} a = \log_{10}\left(\frac{b}{a}\right).$$

(2) Use the definition of β to get:

Equation:

$$\beta_2 - \beta_1 = 10 \log_{10}\left(\frac{I_2}{I_1}\right) = 10 \log_{10} 2.00 = 10\ (0.301)\ \text{dB}.$$

Thus,

Equation:

$$\beta_2 - \beta_1 = 3.01\ \text{dB}.$$

Discussion

This means that the two sound intensity levels differ by 3.01 dB, or about 3 dB, as advertised. Note that because only the ratio I_2/I_1 is given (and not the actual intensities), this result is true for any intensities that differ by a factor of two. For example, a 56.0 dB sound is twice as intense as a 53.0 dB sound, a 97.0 dB sound is half as intense as a 100 dB sound, and so on.

It should be noted at this point that there is another decibel scale in use, called the **sound pressure level**, based on the ratio of the pressure amplitude to a reference pressure. This scale is used particularly in applications where sound travels in water. It is beyond the scope of most introductory texts to treat this scale because it is not commonly used for sounds in air, but it is important to note that very different decibel levels may be encountered when sound pressure levels are quoted. For example, ocean noise pollution produced by ships may be as great as 200 dB expressed in the sound pressure level, where the more familiar sound intensity level we use here would be something under 140 dB for the same sound.

Note:

Take-Home Investigation: Feeling Sound

Find a CD player and a CD that has rock music. Place the player on a light table, insert the CD into the player, and start playing the CD. Place your hand gently on the table next to the speakers. Increase the volume and note the level when the table just begins to vibrate as the rock music plays. Increase the reading on the volume control until it doubles. What has happened to the vibrations?

Exercise:
Check Your Understanding

Problem:

Describe how amplitude is related to the loudness of a sound.

Solution:

Amplitude is directly proportional to the experience of loudness. As amplitude increases, loudness increases.

Exercise:
Check Your Understanding

Problem:

Identify common sounds at the levels of 10 dB, 50 dB, and 100 dB.

Solution:

10 dB: Running fingers through your hair.

50 dB: Inside a quiet home with no television or radio.

100 dB: Take-off of a jet plane.

Section Summary

- Intensity is the same for a sound wave as was defined for all waves; it is
  Equation:

  $$I = \frac{P}{A},$$

  where P is the power crossing area A. The SI unit for I is watts per meter squared. The intensity of a sound wave is also related to the pressure amplitude Δp
  Equation:

  $$I = \frac{(\Delta p)^2}{2\rho v_{\text{w}}},$$

  where ρ is the density of the medium in which the sound wave travels and v_{w} is the speed of sound in the medium.

- Sound intensity level in units of decibels (dB) is
  Equation:

  $$\beta\,(\text{dB}) = 10\,\log_{10}\left(\frac{I}{I_0}\right),$$

  where $I_0 = 10^{-12}\ \text{W/m}^2$ is the threshold intensity of hearing.

Conceptual Questions

Exercise:

Problem:

Six members of a synchronized swim team wear earplugs to protect themselves against water pressure at depths, but they can still hear the music and perform the combinations in the water perfectly. One day, they were asked to leave the pool so the dive team could practice a few dives, and they tried to practice on a mat, but seemed to have a lot more difficulty. Why might this be?

Exercise:

Problem:

A community is concerned about a plan to bring train service to their downtown from the town's outskirts. The current sound intensity level, even though the rail yard is blocks away, is 70 dB downtown. The mayor assures the public that there will be a difference of only 30 dB in sound in the downtown area. Should the townspeople be concerned? Why?

Problems & Exercises

Exercise:

Problem:

What is the intensity in watts per meter squared of 85.0-dB sound?

Solution:
Equation:

$$3.16 \times 10^{-4}\ \mathrm{W/m^2}$$

Exercise:

Problem:

The warning tag on a lawn mower states that it produces noise at a level of 91.0 dB. What is this in watts per meter squared?

Exercise:

Problem:

A sound wave traveling in $20^{\circ}\mathrm{C}$ air has a pressure amplitude of 0.5 Pa. What is the intensity of the wave?

Solution:
Equation:

$$3.04 \times 10^{-4}\ \mathrm{W/m^2}$$

Exercise:

Problem:

What intensity level does the sound in the preceding problem correspond to?

Exercise:

Problem:

What sound intensity level in dB is produced by earphones that create an intensity of $4.00 \times 10^{-2}\ \mathrm{W/m^2}$?

Solution:

106 dB

Exercise:

Problem:

Show that an intensity of $10^{-12}\ \mathrm{W/m^2}$ is the same as $10^{-16}\ \mathrm{W/cm^2}$.

Exercise:

Problem:

(a) What is the decibel level of a sound that is twice as intense as a 90.0-dB sound? (b) What is the decibel level of a sound that is one-fifth as intense as a 90.0-dB sound?

Solution:

(a) 93 dB

(b) 83 dB

Exercise:

Problem:

(a) What is the intensity of a sound that has a level 7.00 dB lower than a $4.00 \times 10^{-9}\ \mathrm{W/m^2}$ sound? (b) What is the intensity of a sound that is 3.00 dB higher than a $4.00 \times 10^{-9}\ \mathrm{W/m^2}$ sound?

Exercise:

Problem:

(a) How much more intense is a sound that has a level 17.0 dB higher than another? (b) If one sound has a level 23.0 dB less than another, what is the ratio of their intensities?

Solution:

(a) 50.1

(b) 5.01×10^{-3} or $\frac{1}{200}$

Exercise:

Problem:

People with good hearing can perceive sounds as low in level as –8.00 dB at a frequency of 3000 Hz. What is the intensity of this sound in watts per meter squared?

Exercise:

Problem:

If a large housefly 3.0 m away from you makes a noise of 40.0 dB, what is the noise level of 1000 flies at that distance, assuming interference has a negligible effect?

Solution:

70.0 dB

Exercise:

Problem:

Ten cars in a circle at a boom box competition produce a 120-dB sound intensity level at the center of the circle. What is the average sound intensity level produced there by each stereo, assuming interference effects can be neglected?

Exercise:

Problem:

The amplitude of a sound wave is measured in terms of its maximum gauge pressure. By what factor does the amplitude of a sound wave increase if the sound intensity level goes up by 40.0 dB?

Solution:

100

Exercise:

Problem:

If a sound intensity level of 0 dB at 1000 Hz corresponds to a maximum gauge pressure (sound amplitude) of $10^{-9}\ \text{atm}$, what is the maximum gauge pressure in a 60-dB sound? What is the maximum gauge pressure in a 120-dB sound?

Exercise:

Problem:

An 8-hour exposure to a sound intensity level of 90.0 dB may cause hearing damage. What energy in joules falls on a 0.800-cm-diameter eardrum so exposed?

Solution:
Equation:

$$1.45 \times 10^{-3}\ \text{J}$$

Exercise:

Problem:

(a) Ear trumpets were never very common, but they did aid people with hearing losses by gathering sound over a large area and concentrating it on the smaller area of the eardrum. What decibel increase does an ear trumpet produce if its sound gathering area is $900\ \text{cm}^2$ and the area of the eardrum is $0.500\ \text{cm}^2$, but the trumpet only has an efficiency of 5.00% in transmitting the sound to the eardrum? (b) Comment on the usefulness of the decibel increase found in part (a).

Exercise:

Problem:

Sound is more effectively transmitted into a stethoscope by direct contact than through the air, and it is further intensified by being concentrated on the smaller area of the eardrum. It is reasonable to assume that sound is transmitted into a stethoscope 100 times as effectively compared with transmission though the air. What, then, is the gain in decibels produced by a stethoscope that has a sound gathering area of 15.0 cm^2, and concentrates the sound onto two eardrums with a total area of 0.900 cm^2 with an efficiency of 40.0%?

Solution:

28.2 dB

Exercise:

Problem:

Loudspeakers can produce intense sounds with surprisingly small energy input in spite of their low efficiencies. Calculate the power input needed to produce a 90.0-dB sound intensity level for a 12.0-cm-diameter speaker that has an efficiency of 1.00%. (This value is the sound intensity level right at the speaker.)

Glossary

intensity
: the power per unit area carried by a wave

sound intensity level
: a unitless quantity telling you the level of the sound relative to a fixed standard

sound pressure level
: the ratio of the pressure amplitude to a reference pressure

Doppler Effect and Sonic Booms

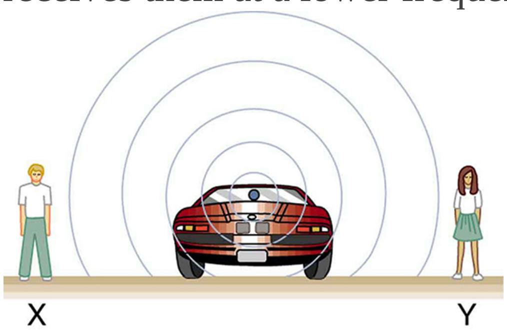
- Define Doppler effect, Doppler shift, and sonic boom.
- Calculate the frequency of a sound heard by someone observing Doppler shift.
- Describe the sounds produced by objects moving faster than the speed of sound.

The characteristic sound of a motorcycle buzzing by is an example of the **Doppler effect**. The high-pitch scream shifts dramatically to a lower-pitch roar as the motorcycle passes by a stationary observer. The closer the motorcycle brushes by, the more abrupt the shift. The faster the motorcycle moves, the greater the shift. We also hear this characteristic shift in frequency for passing race cars, airplanes, and trains. It is so familiar that it is used to imply motion and children often mimic it in play.

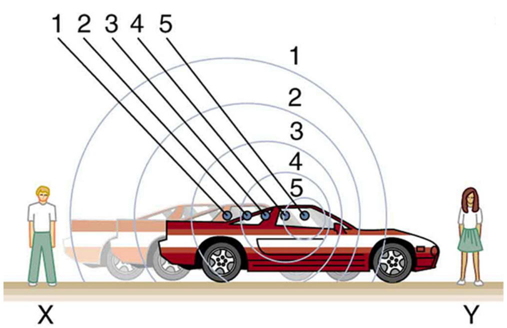
The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer. Although less familiar, this effect is easily noticed for a stationary source and moving observer. For example, if you ride a train past a stationary warning bell, you will hear the bell's frequency shift from high to low as you pass by. The actual change in frequency due to relative motion of source and observer is called a **Doppler shift**. The Doppler effect and Doppler shift are named for the Austrian physicist and mathematician Christian Johann Doppler (1803–1853), who did experiments with both moving sources and moving observers. Doppler, for example, had musicians play on a moving open train car and also play standing next to the train tracks as a train passed by. Their music was observed both on and off the train, and changes in frequency were measured.

What causes the Doppler shift? [link], [link], and [link] compare sound waves emitted by stationary and moving sources in a stationary air mass. Each disturbance spreads out spherically from the point where the sound was emitted. If the source is stationary, then all of the spheres representing the air compressions in the sound wave centered on the same point, and the stationary observers on either side see the same wavelength and frequency as emitted by the source, as in [link]. If the source is moving, as in [link], then the situation is different. Each compression of the air moves out in a

sphere from the point where it was emitted, but the point of emission moves. This moving emission point causes the air compressions to be closer together on one side and farther apart on the other. Thus, the wavelength is shorter in the direction the source is moving (on the right in [link]), and longer in the opposite direction (on the left in [link]). Finally, if the observers move, as in [link], the frequency at which they receive the compressions changes. The observer moving toward the source receives them at a higher frequency, and the person moving away from the source receives them at a lower frequency.

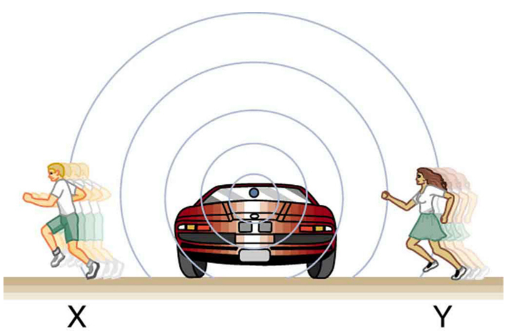


Sounds emitted by a source spread out in spherical waves. Because the source, observers, and air are stationary, the wavelength and frequency are the same in all directions and to all observers.



Sounds emitted by a

source moving to the right spread out from the points at which they were emitted. The wavelength is reduced and, consequently, the frequency is increased in the direction of motion, so that the observer on the right hears a higher-pitch sound. The opposite is true for the observer on the left, where the wavelength is increased and the frequency is reduced.



The same effect is produced when the observers move relative to the source. Motion toward the source increases frequency as the observer on the right passes through more wave crests than she would if stationary. Motion away from the

source decreases frequency as the observer on the left passes through fewer wave crests than he would if stationary.

We know that wavelength and frequency are related by $v_{\text{w}} = f\lambda$, where v_{w} is the fixed speed of sound. The sound moves in a medium and has the same speed v_{w} in that medium whether the source is moving or not. Thus f multiplied by λ is a constant. Because the observer on the right in [link] receives a shorter wavelength, the frequency she receives must be higher. Similarly, the observer on the left receives a longer wavelength, and hence he hears a lower frequency. The same thing happens in [link]. A higher frequency is received by the observer moving toward the source, and a lower frequency is received by an observer moving away from the source. In general, then, relative motion of source and observer toward one another increases the received frequency. Relative motion apart decreases frequency. The greater the relative speed is, the greater the effect.

Note:
The Doppler Effect
The Doppler effect occurs not only for sound but for any wave when there is relative motion between the observer and the source. There are Doppler shifts in the frequency of sound, light, and water waves, for example. Doppler shifts can be used to determine velocity, such as when ultrasound is reflected from blood in a medical diagnostic. The recession of galaxies is determined by the shift in the frequencies of light received from them and has implied much about the origins of the universe. Modern physics has been profoundly affected by observations of Doppler shifts.

For a stationary observer and a moving source, the frequency f_{obs} received by the observer can be shown to be

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right),$$

where f_{s} is the frequency of the source, v_{s} is the speed of the source along a line joining the source and observer, and v_{w} is the speed of sound. The minus sign is used for motion toward the observer and the plus sign for motion away from the observer, producing the appropriate shifts up and down in frequency. Note that the greater the speed of the source, the greater the effect. Similarly, for a stationary source and moving observer, the frequency received by the observer f_{obs} is given by

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right),$$

where v_{obs} is the speed of the observer along a line joining the source and observer. Here the plus sign is for motion toward the source, and the minus is for motion away from the source.

Example:

Calculate Doppler Shift: A Train Horn

Suppose a train that has a 150-Hz horn is moving at 35.0 m/s in still air on a day when the speed of sound is 340 m/s.

(a) What frequencies are observed by a stationary person at the side of the tracks as the train approaches and after it passes?

(b) What frequency is observed by the train’s engineer traveling on the train?

Strategy

To find the observed frequency in (a), $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right)$, must be used because the source is moving. The minus sign is used for the approaching

train, and the plus sign for the receding train. In (b), there are two Doppler shifts—one for a moving source and the other for a moving observer.

Solution for (a)

(1) Enter known values into $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{s}}}\right)$.

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{s}}}\right) = (150\text{ Hz})\left(\frac{340\text{ m/s}}{340\text{ m/s} - 35.0\text{ m/s}}\right)$$

(2) Calculate the frequency observed by a stationary person as the train approaches.

Equation:

$$f_{\text{obs}} = (150\text{ Hz})(1.11) = 167\text{ Hz}$$

(3) Use the same equation with the plus sign to find the frequency heard by a stationary person as the train recedes.

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} + v_{\text{s}}}\right) = (150\text{ Hz})\left(\frac{340\text{ m/s}}{340\text{ m/s} + 35.0\text{ m/s}}\right)$$

(4) Calculate the second frequency.

Equation:

$$f_{\text{obs}} = (150\text{ Hz})(0.907) = 136\text{ Hz}$$

Discussion on (a)

The numbers calculated are valid when the train is far enough away that the motion is nearly along the line joining train and observer. In both cases, the shift is significant and easily noticed. Note that the shift is 17.0 Hz for motion toward and 14.0 Hz for motion away. The shifts are not symmetric.

Solution for (b)

(1) Identify knowns:

- It seems reasonable that the engineer would receive the same frequency as emitted by the horn, because the relative velocity

between them is zero.

- Relative to the medium (air), the speeds are $v_{\text{s}} = v_{\text{obs}} = 35.0 \text{ m/s}$.
- The first Doppler shift is for the moving observer; the second is for the moving source.

(2) Use the following equation:
Equation:

$$f_{\text{obs}} = \left[f_{\text{s}} \left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}} \right) \right] \left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}} \right).$$

The quantity in the square brackets is the Doppler-shifted frequency due to a moving observer. The factor on the right is the effect of the moving source.
(3) Because the train engineer is moving in the direction toward the horn, we must use the plus sign for v_{obs}; however, because the horn is also moving in the direction away from the engineer, we also use the plus sign for v_{s}. But the train is carrying both the engineer and the horn at the same velocity, so $v_{\text{s}} = v_{\text{obs}}$. As a result, everything but f_{s} cancels, yielding
Equation:

$$f_{\text{obs}}=f_{\text{s}}.$$

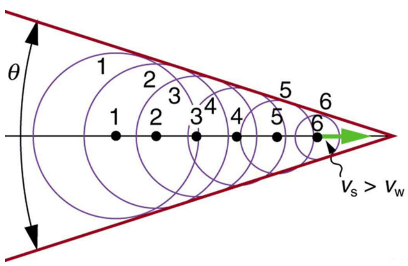
Discussion for (b)
We may expect that there is no change in frequency when source and observer move together because it fits your experience. For example, there is no Doppler shift in the frequency of conversations between driver and passenger on a motorcycle. People talking when a wind moves the air between them also observe no Doppler shift in their conversation. The crucial point is that source and observer are not moving relative to each other.

Sonic Booms to Bow Wakes

What happens to the sound produced by a moving source, such as a jet airplane, that approaches or even exceeds the speed of sound? The answer

to this question applies not only to sound but to all other waves as well.

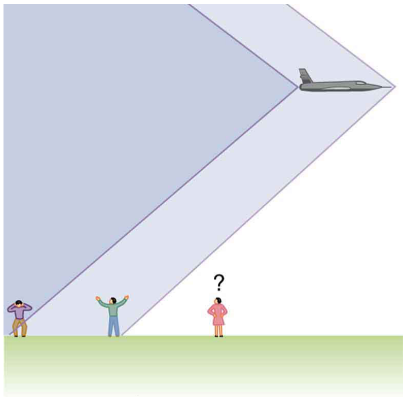
Suppose a jet airplane is coming nearly straight at you, emitting a sound of frequency f_{s}. The greater the plane's speed v_{s}, the greater the Doppler shift and the greater the value observed for f_{obs}. Now, as v_{s} approaches the speed of sound, f_{obs} approaches infinity, because the denominator in $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right)$ approaches zero. At the speed of sound, this result means that in front of the source, each successive wave is superimposed on the previous one because the source moves forward at the speed of sound. The observer gets them all at the same instant, and so the frequency is infinite. (Before airplanes exceeded the speed of sound, some people argued it would be impossible because such constructive superposition would produce pressures great enough to destroy the airplane.) If the source exceeds the speed of sound, no sound is received by the observer until the source has passed, so that the sounds from the approaching source are mixed with those from it when receding. This mixing appears messy, but something interesting happens—a sonic boom is created. (See [link].)



Sound waves from a source that moves faster than the speed of sound spread spherically from the point where they are emitted, but the source moves ahead of each.

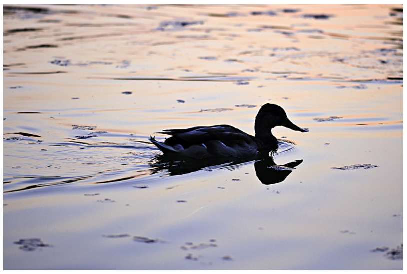
Constructive interference along the lines shown (actually a cone in three dimensions) creates a shock wave called a sonic boom. The faster the speed of the source, the smaller the angle θ.

There is constructive interference along the lines shown (a cone in three dimensions) from similar sound waves arriving there simultaneously. This superposition forms a disturbance called a **sonic boom**, a constructive interference of sound created by an object moving faster than sound. Inside the cone, the interference is mostly destructive, and so the sound intensity there is much less than on the shock wave. An aircraft creates two sonic booms, one from its nose and one from its tail. (See [link].) During television coverage of space shuttle landings, two distinct booms could often be heard. These were separated by exactly the time it would take the shuttle to pass by a point. Observers on the ground often do not see the aircraft creating the sonic boom, because it has passed by before the shock wave reaches them, as seen in [link]. If the aircraft flies close by at low altitude, pressures in the sonic boom can be destructive and break windows as well as rattle nerves. Because of how destructive sonic booms can be, supersonic flights are banned over populated areas of the United States.

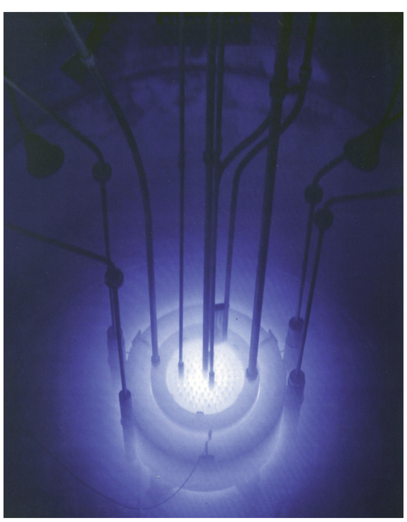


Two sonic booms, created by the nose and tail of an aircraft, are observed on the ground after the plane has passed by.

Sonic booms are one example of a broader phenomenon called bow wakes. A **bow wake**, such as the one in [link], is created when the wave source moves faster than the wave propagation speed. Water waves spread out in circles from the point where created, and the bow wake is the familiar V-shaped wake trailing the source. A more exotic bow wake is created when a subatomic particle travels through a medium faster than the speed of light travels in that medium. (In a vacuum, the maximum speed of light will be $c = 3.00 \times 10^8 \text{ m/s}$; in the medium of water, the speed of light is closer to $0.75c$. If the particle creates light in its passage, that light spreads on a cone with an angle indicative of the speed of the particle, as illustrated in [link]. Such a bow wake is called Cerenkov radiation and is commonly observed in particle physics.



Bow wake created by a duck. Constructive interference produces the rather structured wake, while there is relatively little wave action inside the wake, where interference is mostly destructive. (credit: Horia Varlan, Flickr)



The blue glow in this research reactor pool is Cerenkov radiation caused by subatomic particles traveling faster than the speed of light in water. (credit: U.S. Nuclear Regulatory Commission)

Doppler shifts and sonic booms are interesting sound phenomena that occur in all types of waves. They can be of considerable use. For example, the Doppler shift in ultrasound can be used to measure blood velocity, while police use the Doppler shift in radar (a microwave) to measure car velocities. In meteorology, the Doppler shift is used to track the motion of storm clouds; such “Doppler Radar” can give velocity and direction and rain or snow potential of imposing weather fronts. In astronomy, we can examine the light emitted from distant galaxies and determine their speed relative to ours. As galaxies move away from us, their light is shifted to a lower frequency, and so to a longer wavelength—the so-called red shift. Such information from galaxies far, far away has allowed us to estimate the age of the universe (from the Big Bang) as about 14 billion years.

Exercise:

Check Your Understanding

Problem:

Why did scientist Christian Doppler observe musicians both on a moving train and also from a stationary point not on the train?

Solution:

Doppler needed to compare the perception of sound when the observer is stationary and the sound source moves, as well as when the sound

source and the observer are both in motion.

Exercise:
Check Your Understanding

Problem:

Describe a situation in your life when you might rely on the Doppler shift to help you either while driving a car or walking near traffic.

Solution:

If I am driving and I hear Doppler shift in an ambulance siren, I would be able to tell when it was getting closer and also if it has passed by. This would help me to know whether I needed to pull over and let the ambulance through.

Section Summary

- The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer.
- The actual change in frequency is called the Doppler shift.
- A sonic boom is constructive interference of sound created by an object moving faster than sound.
- A sonic boom is a type of bow wake created when any wave source moves faster than the wave propagation speed.
- For a stationary observer and a moving source, the observed frequency f_{obs} is:
  Equation:
  $$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right),$$
  where f_{s} is the frequency of the source, v_{s} is the speed of the source, and v_{w} is the speed of sound. The minus sign is used for motion toward the observer and the plus sign for motion away.
- For a stationary source and moving observer, the observed frequency is:

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right),$$

where v_{obs} is the speed of the observer.

Conceptual Questions

Exercise:

Problem: Is the Doppler shift real or just a sensory illusion?

Exercise:

Problem:

Due to efficiency considerations related to its bow wake, the supersonic transport aircraft must maintain a cruising speed that is a constant ratio to the speed of sound (a constant Mach number). If the aircraft flies from warm air into colder air, should it increase or decrease its speed? Explain your answer.

Exercise:

Problem:

When you hear a sonic boom, you often cannot see the plane that made it. Why is that?

Problems & Exercises

Exercise:

Problem:

(a) What frequency is received by a person watching an oncoming ambulance moving at 110 km/h and emitting a steady 800-Hz sound from its siren? The speed of sound on this day is 345 m/s. (b) What frequency does she receive after the ambulance has passed?

Solution:

(a) 878 Hz

(b) 735 Hz

Exercise:

Problem:

(a) At an air show a jet flies directly toward the stands at a speed of 1200 km/h, emitting a frequency of 3500 Hz, on a day when the speed of sound is 342 m/s. What frequency is received by the observers? (b) What frequency do they receive as the plane flies directly away from them?

Exercise:

Problem:

What frequency is received by a mouse just before being dispatched by a hawk flying at it at 25.0 m/s and emitting a screech of frequency 3500 Hz? Take the speed of sound to be 331 m/s.

Solution:
Equation:

$$3.79 \times 10^3 \text{ Hz}$$

Exercise:

Problem:

A spectator at a parade receives an 888-Hz tone from an oncoming trumpeter who is playing an 880-Hz note. At what speed is the musician approaching if the speed of sound is 338 m/s?

Exercise:

Problem:

A commuter train blows its 200-Hz horn as it approaches a crossing. The speed of sound is 335 m/s. (a) An observer waiting at the crossing receives a frequency of 208 Hz. What is the speed of the train? (b) What frequency does the observer receive as the train moves away?

Solution:

(a) 12.9 m/s

(b) 193 Hz

Exercise:

Problem:

Can you perceive the shift in frequency produced when you pull a tuning fork toward you at 10.0 m/s on a day when the speed of sound is 344 m/s? To answer this question, calculate the factor by which the frequency shifts and see if it is greater than 0.300%.

Exercise:

Problem:

Two eagles fly directly toward one another, the first at 15.0 m/s and the second at 20.0 m/s. Both screech, the first one emitting a frequency of 3200 Hz and the second one emitting a frequency of 3800 Hz. What frequencies do they receive if the speed of sound is 330 m/s?

Solution:

First eagle hears 4.23×10^3 Hz

Second eagle hears 3.56×10^3 Hz

Exercise:

Problem:

What is the minimum speed at which a source must travel toward you for you to be able to hear that its frequency is Doppler shifted? That is, what speed produces a shift of 0.300% on a day when the speed of sound is 331 m/s?

Glossary

Doppler effect
: an alteration in the observed frequency of a sound due to motion of either the source or the observer

Doppler shift
: the actual change in frequency due to relative motion of source and observer

sonic boom
: a constructive interference of sound created by an object moving faster than sound

bow wake
: V-shaped disturbance created when the wave source moves faster than the wave propagation speed

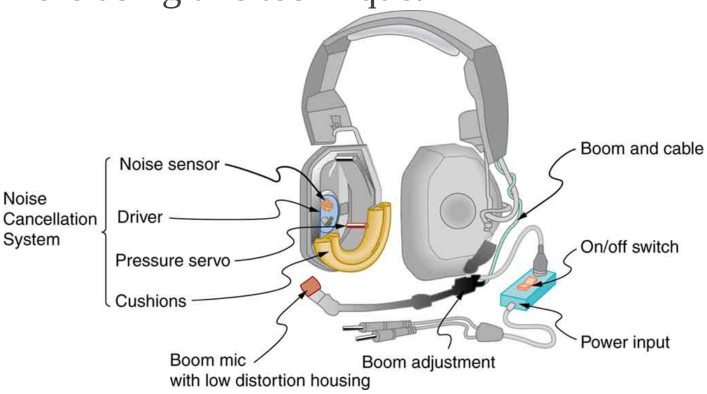
Sound Interference and Resonance: Standing Waves in Air Columns

- Define antinode, node, fundamental, overtones, and harmonics.
- Identify instances of sound interference in everyday situations.
- Describe how sound interference occurring inside open and closed tubes changes the characteristics of the sound, and how this applies to sounds produced by musical instruments.
- Calculate the length of a tube using sound wave measurements.

Some types of headphones use the phenomena of constructive and destructive interference to cancel out outside noises. (credit: JVC America, Flickr)

Interference is the hallmark of waves, all of which exhibit constructive and destructive interference exactly analogous to that seen for water waves. In fact, one way to prove something "is a wave" is to observe interference effects. So, sound being a wave, we expect it to exhibit interference; we have already mentioned a few such effects, such as the beats from two similar notes played simultaneously.

[link] shows a clever use of sound interference to cancel noise. Larger-scale applications of active noise reduction by destructive interference are contemplated for entire passenger compartments in commercial aircraft. To obtain destructive interference, a fast electronic analysis is performed, and a second sound is introduced with its maxima and minima exactly reversed from the incoming noise. Sound waves in fluids are pressure waves and consistent with Pascal's principle; pressures from two different sources add and subtract like simple numbers; that is, positive and negative gauge pressures add to a much smaller pressure, producing a lower-intensity sound. Although completely destructive interference is possible only under the simplest conditions, it is possible to reduce noise levels by 30 dB or more using this technique.



Headphones designed to cancel noise with destructive interference create a sound wave exactly opposite to the incoming sound. These headphones can be more effective than the simple passive attenuation used in most ear protection. Such headphones were

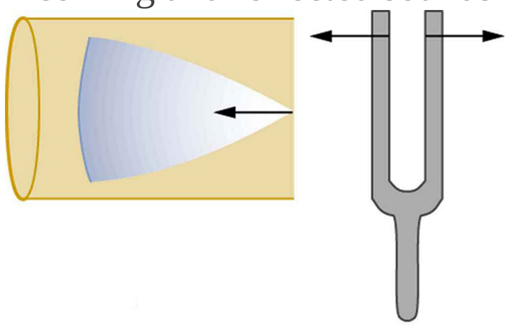
used on the record-setting, around the world nonstop flight of the Voyager aircraft to protect the pilots' hearing from engine noise.

Where else can we observe sound interference? All sound resonances, such as in musical instruments, are due to constructive and destructive interference. Only the resonant frequencies interfere constructively to form standing waves, while others interfere destructively and are absent. From the toot made by blowing over a bottle, to the characteristic flavor of a violin's sounding box, to the recognizability of a great singer's voice, resonance and standing waves play a vital role.

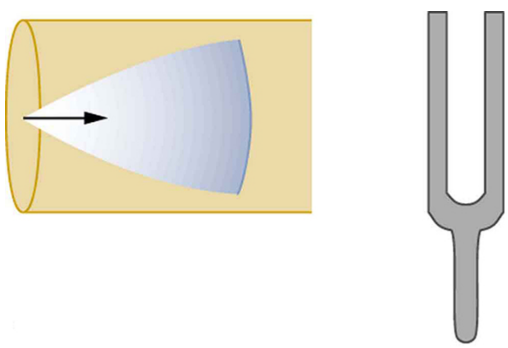
Note:
Interference
Interference is such a fundamental aspect of waves that observing interference is proof that something is a wave. The wave nature of light was established by experiments showing interference. Similarly, when electrons scattered from crystals exhibited interference, their wave nature was confirmed to be exactly as predicted by symmetry with certain wave characteristics of light.

Suppose we hold a tuning fork near the end of a tube that is closed at the other end, as shown in [link], [link], [link], and [link]. If the tuning fork has just the right frequency, the air column in the tube resonates loudly, but at most frequencies it vibrates very little. This observation just means that the air column has only certain natural frequencies. The figures show how a resonance at the lowest of these natural frequencies is formed. A disturbance travels down the tube at the speed of sound and bounces off the closed end. If the tube is just the right length, the reflected sound arrives back at the tuning fork exactly half a cycle later, and it interferes

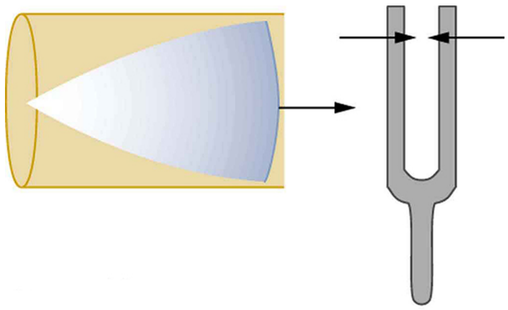
constructively with the continuing sound produced by the tuning fork. The incoming and reflected sounds form a standing wave in the tube as shown.



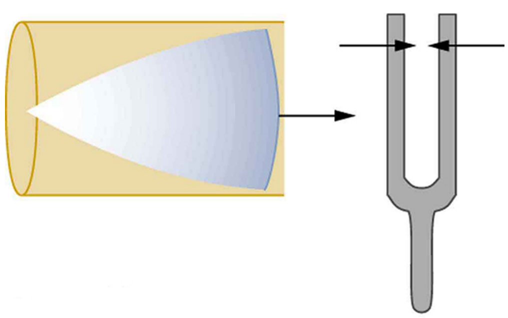
Resonance of air in a tube closed at one end, caused by a tuning fork. A disturbance moves down the tube.



Resonance of air in a tube closed at one end, caused by a tuning fork. The disturbance reflects from the closed end of the tube.



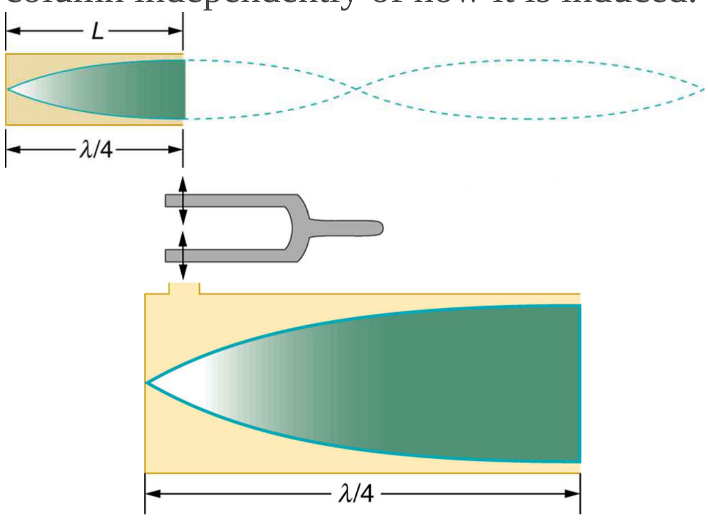
Resonance of air in a tube closed at one end, caused by a tuning fork. If the length of the tube L is just right, the disturbance gets back to the tuning fork half a cycle later and interferes constructively with the continuing sound from the tuning fork. This interference forms a standing wave, and the air column resonates.



Resonance of air in a tube closed at one end, caused by a tuning fork. A graph of air displacement along the length of the tube shows none at the closed

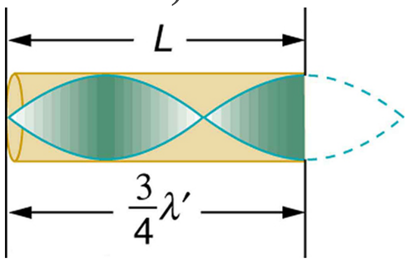
end, where the motion is constrained, and a maximum at the open end. This standing wave has one-fourth of its wavelength in the tube, so that $\lambda = 4L$.

The standing wave formed in the tube has its maximum air displacement (an **antinode**) at the open end, where motion is unconstrained, and no displacement (a **node**) at the closed end, where air movement is halted. The distance from a node to an antinode is one-fourth of a wavelength, and this equals the length of the tube; thus, $\lambda = 4L$. This same resonance can be produced by a vibration introduced at or near the closed end of the tube, as shown in [link]. It is best to consider this a natural vibration of the air column independently of how it is induced.

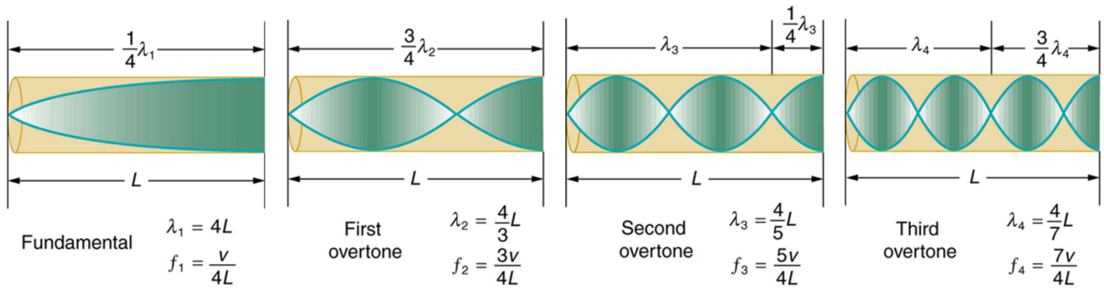


The same standing wave is created in the tube by a vibration introduced near its closed end.

Given that maximum air displacements are possible at the open end and none at the closed end, there are other, shorter wavelengths that can resonate in the tube, such as the one shown in [link]. Here the standing wave has three-fourths of its wavelength in the tube, or $L = (3/4)\lambda\prime$, so that $\lambda\prime = 4L/3$. Continuing this process reveals a whole series of shorter-wavelength and higher-frequency sounds that resonate in the tube. We use specific terms for the resonances in any system. The lowest resonant frequency is called the **fundamental**, while all higher resonant frequencies are called **overtones**. All resonant frequencies are integral multiples of the fundamental, and they are collectively called **harmonics**. The fundamental is the first harmonic, the first overtone is the second harmonic, and so on. [link] shows the fundamental and the first three overtones (the first four harmonics) in a tube closed at one end.

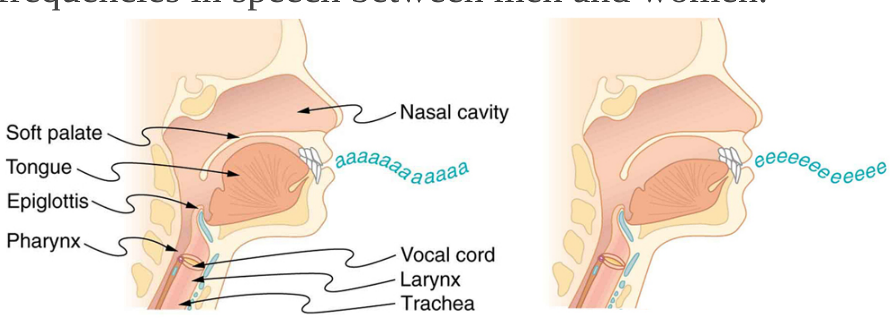


Another resonance for a tube closed at one end. This has maximum air displacements at the open end, and none at the closed end. The wavelength is shorter, with three-fourths $\lambda\prime$ equaling the length of the tube, so that $\lambda\prime = 4L/3$. This higher-frequency vibration is the first overtone.



The fundamental and three lowest overtones for a tube closed at one end. All have maximum air displacements at the open end and none at the closed end.

The fundamental and overtones can be present simultaneously in a variety of combinations. For example, middle C on a trumpet has a sound distinctively different from middle C on a clarinet, both instruments being modified versions of a tube closed at one end. The fundamental frequency is the same (and usually the most intense), but the overtones and their mix of intensities are different and subject to shading by the musician. This mix is what gives various musical instruments (and human voices) their distinctive characteristics, whether they have air columns, strings, sounding boxes, or drumheads. In fact, much of our speech is determined by shaping the cavity formed by the throat and mouth and positioning the tongue to adjust the fundamental and combination of overtones. Simple resonant cavities can be made to resonate with the sound of the vowels, for example. (See [link].) In boys, at puberty, the larynx grows and the shape of the resonant cavity changes giving rise to the difference in predominant frequencies in speech between men and women.



The throat and mouth form an air column closed at one end that resonates in response to vibrations in the voice box. The spectrum of overtones and their intensities vary with mouth shaping and tongue position to form different sounds. The voice box can be replaced with a mechanical vibrator, and understandable speech is still possible. Variations in basic shapes make different voices recognizable.

Now let us look for a pattern in the resonant frequencies for a simple tube that is closed at one end. The fundamental has $\lambda = 4L$, and frequency is related to wavelength and the speed of sound as given by:
Equation:

$$v_{\mathrm{w}} = f\lambda.$$

Solving for f in this equation gives
Equation:

$$f = \frac{v_{\mathrm{w}}}{\lambda} = \frac{v_{\mathrm{w}}}{4L},$$

where v_{w} is the speed of sound in air. Similarly, the first overtone has $\lambda\prime = 4L/3$ (see [link]), so that
Equation:

$$f\prime = 3\frac{v_{\mathrm{w}}}{4L} = 3f.$$

Because $f\prime = 3f$, we call the first overtone the third harmonic. Continuing this process, we see a pattern that can be generalized in a single expression. The resonant frequencies of a tube closed at one end are
Equation:

$$f_n = n\frac{v_w}{4L}, n = 1,3,5,$$

where f_1 is the fundamental, f_3 is the first overtone, and so on. It is interesting that the resonant frequencies depend on the speed of sound and, hence, on temperature. This dependence poses a noticeable problem for organs in old unheated cathedrals, and it is also the reason why musicians commonly bring their wind instruments to room temperature before playing them.

Example:
Find the Length of a Tube with a 128 Hz Fundamental
(a) What length should a tube closed at one end have on a day when the air temperature, is $22.0^{\circ}C$, if its fundamental frequency is to be 128 Hz (C below middle C)?
(b) What is the frequency of its fourth overtone?
Strategy
The length L can be found from the relationship in $f_n = n\frac{v_w}{4L}$, but we will first need to find the speed of sound v_w.
Solution for (a)
(1) Identify knowns:

- the fundamental frequency is 128 Hz
- the air temperature is $22.0^{\circ}C$

(2) Use $f_n = n\frac{v_w}{4L}$ to find the fundamental frequency ($n = 1$).
Equation:

$$f_1 = \frac{v_w}{4L}$$

(3) Solve this equation for length.
Equation:

$$L = \frac{v_w}{4f_1}$$

(4) Find the speed of sound using $v_{\mathrm{w}} = (331\ \mathrm{m/s})\sqrt{\frac{T}{273\ \mathrm{K}}}$.

Equation:

$$v_{\mathrm{w}} = (331\ \mathrm{m/s})\sqrt{\frac{295\ \mathrm{K}}{273\ \mathrm{K}}} = 344\ \mathrm{m/s}$$

(5) Enter the values of the speed of sound and frequency into the expression for L.

Equation:

$$L = \frac{v_{\mathrm{w}}}{4f_1} = \frac{344\ \mathrm{m/s}}{4(128\ \mathrm{Hz})} = 0.672\ \mathrm{m}$$

Discussion on (a)

Many wind instruments are modified tubes that have finger holes, valves, and other devices for changing the length of the resonating air column and hence, the frequency of the note played. Horns producing very low frequencies, such as tubas, require tubes so long that they are coiled into loops.

Solution for (b)

(1) Identify knowns:

- the first overtone has $n = 3$
- the second overtone has $n = 5$
- the third overtone has $n = 7$
- the fourth overtone has $n = 9$

(2) Enter the value for the fourth overtone into $f_n = n\frac{v_{\mathrm{w}}}{4L}$.

Equation:

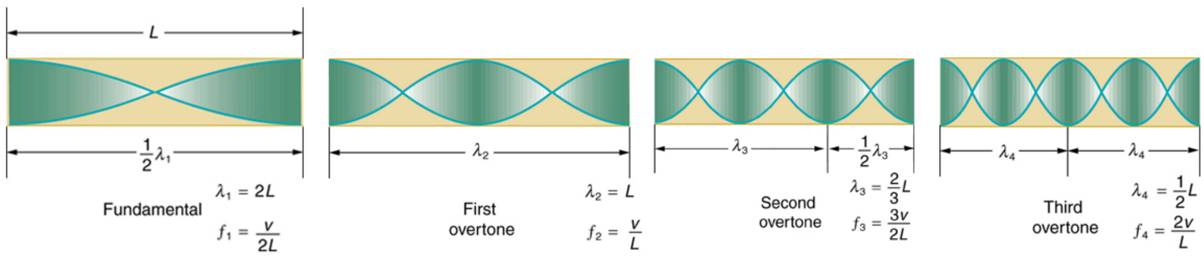
$$f_9 = 9\frac{v_{\mathrm{w}}}{4L} = 9f_1 = 1.15\ \mathrm{kHz}$$

Discussion on (b)

Whether this overtone occurs in a simple tube or a musical instrument depends on how it is stimulated to vibrate and the details of its shape. The

trombone, for example, does not produce its fundamental frequency and only makes overtones.

Another type of tube is one that is *open* at both ends. Examples are some organ pipes, flutes, and oboes. The resonances of tubes open at both ends can be analyzed in a very similar fashion to those for tubes closed at one end. The air columns in tubes open at both ends have maximum air displacements at both ends, as illustrated in [link]. Standing waves form as shown.



The resonant frequencies of a tube open at both ends are shown, including the fundamental and the first three overtones. In all cases the maximum air displacements occur at both ends of the tube, giving it different natural frequencies than a tube closed at one end.

Based on the fact that a tube open at both ends has maximum air displacements at both ends, and using [link] as a guide, we can see that the resonant frequencies of a tube open at both ends are:

Equation:

$$f_n = n\frac{v_{\text{w}}}{2L},\ n = 1, 2, 3...,$$

where f_1 is the fundamental, f_2 is the first overtone, f_3 is the second overtone, and so on. Note that a tube open at both ends has a fundamental frequency twice what it would have if closed at one end. It also has a different spectrum of overtones than a tube closed at one end. So if you had

two tubes with the same fundamental frequency but one was open at both ends and the other was closed at one end, they would sound different when played because they have different overtones. Middle C, for example, would sound richer played on an open tube, because it has even multiples of the fundamental as well as odd. A closed tube has only odd multiples.

Note:
Real-World Applications: Resonance in Everyday Systems
Resonance occurs in many different systems, including strings, air columns, and atoms. Resonance is the driven or forced oscillation of a system at its natural frequency. At resonance, energy is transferred rapidly to the oscillating system, and the amplitude of its oscillations grows until the system can no longer be described by Hooke’s law. An example of this is the distorted sound intentionally produced in certain types of rock music.

Wind instruments use resonance in air columns to amplify tones made by lips or vibrating reeds. Other instruments also use air resonance in clever ways to amplify sound. [link] shows a violin and a guitar, both of which have sounding boxes but with different shapes, resulting in different overtone structures. The vibrating string creates a sound that resonates in the sounding box, greatly amplifying the sound and creating overtones that give the instrument its characteristic flavor. The more complex the shape of the sounding box, the greater its ability to resonate over a wide range of frequencies. The marimba, like the one shown in [link] uses pots or gourds below the wooden slats to amplify their tones. The resonance of the pot can be adjusted by adding water.

String instruments such as violins and guitars use resonance in their sounding boxes to amplify and enrich the sound created by their vibrating strings. The bridge and supports couple the string vibrations to the sounding boxes and air within. (credits: guitar, Feliciano Guimares, Fotopedia; violin, Steve Snodgrass, Flickr)

Resonance has been used in musical instruments since prehistoric times. This marimba uses gourds as resonance chambers to amplify its sound. (credit: APC Events, Flickr)

We have emphasized sound applications in our discussions of resonance and standing waves, but these ideas apply to any system that has wave characteristics. Vibrating strings, for example, are actually resonating and have fundamentals and overtones similar to those for air columns. More subtle are the resonances in atoms due to the wave character of their electrons. Their orbitals can be viewed as standing waves, which have a fundamental (ground state) and overtones (excited states). It is fascinating that wave characteristics apply to such a wide range of physical systems.

Exercise:

Check Your Understanding

Problem:

Describe how noise-canceling headphones differ from standard headphones used to block outside sounds.

Solution:

Regular headphones only block sound waves with a physical barrier. Noise-canceling headphones use destructive interference to reduce the loudness of outside sounds.

Exercise:
Check Your Understanding

Problem:

How is it possible to use a standing wave's node and antinode to determine the length of a closed-end tube?

Solution:

When the tube resonates at its natural frequency, the wave's node is located at the closed end of the tube, and the antinode is located at the open end. The length of the tube is equal to one-fourth of the wavelength of this wave. Thus, if we know the wavelength of the wave, we can determine the length of the tube.

Note:
PhET Explorations: Sound
This simulation lets you see sound waves. Adjust the frequency or volume and you can see and hear how the wave changes. Move the listener around and hear what she hears.
https://archive.cnx.org/specials/c4d3b96e-41f3-11e5-ab7b-47e22dffc18e/sound/#sim-single-source

Section Summary

- Sound interference and resonance have the same properties as defined for all waves.
- In air columns, the lowest-frequency resonance is called the fundamental, whereas all higher resonant frequencies are called overtones. Collectively, they are called harmonics.

- The resonant frequencies of a tube closed at one end are:
  Equation:

  $$f_n = n\frac{v_w}{4L}, n = 1, 3, 5...,$$

  f_1 is the fundamental and L is the length of the tube.
- The resonant frequencies of a tube open at both ends are:
  Equation:

  $$f_n = n\frac{v_w}{2L}, n = 1, 2, 3...$$

Conceptual Questions

Exercise:

Problem:

How does an unamplified guitar produce sounds so much more intense than those of a plucked string held taut by a simple stick?

Exercise:

Problem:

You are given two wind instruments of identical length. One is open at both ends, whereas the other is closed at one end. Which is able to produce the lowest frequency?

Exercise:

Problem:

What is the difference between an overtone and a harmonic? Are all harmonics overtones? Are all overtones harmonics?

Problems & Exercises

Exercise:

Problem:

A “showy” custom-built car has two brass horns that are supposed to produce the same frequency but actually emit 263.8 and 264.5 Hz. What beat frequency is produced?

Solution:

0.7 Hz

Exercise:

Problem:

What beat frequencies will be present: (a) If the musical notes A and C are played together (frequencies of 220 and 264 Hz)? (b) If D and F are played together (frequencies of 297 and 352 Hz)? (c) If all four are played together?

Exercise:

Problem:

What beat frequencies result if a piano hammer hits three strings that emit frequencies of 127.8, 128.1, and 128.3 Hz?

Solution:

0.3 Hz, 0.2 Hz, 0.5 Hz

Exercise:

Problem:

A piano tuner hears a beat every 2.00 s when listening to a 264.0-Hz tuning fork and a single piano string. What are the two possible frequencies of the string?

Exercise:

Problem:

(a) What is the fundamental frequency of a 0.672-m-long tube, open at both ends, on a day when the speed of sound is 344 m/s? (b) What is the frequency of its second harmonic?

Solution:

(a) 256 Hz

(b) 512 Hz

Exercise:

Problem:

If a wind instrument, such as a tuba, has a fundamental frequency of 32.0 Hz, what are its first three overtones? It is closed at one end. (The overtones of a real tuba are more complex than this example, because it is a tapered tube.)

Exercise:

Problem:

What are the first three overtones of a bassoon that has a fundamental frequency of 90.0 Hz? It is open at both ends. (The overtones of a real bassoon are more complex than this example, because its double reed makes it act more like a tube closed at one end.)

Solution:

180 Hz, 270 Hz, 360 Hz

Exercise:

Problem:

How long must a flute be in order to have a fundamental frequency of 262 Hz (this frequency corresponds to middle C on the evenly tempered chromatic scale) on a day when air temperature is $20.0^{\circ}\mathrm{C}$? It is open at both ends.

Exercise:

Problem:

What length should an oboe have to produce a fundamental frequency of 110 Hz on a day when the speed of sound is 343 m/s? It is open at both ends.

Solution:

1.56 m

Exercise:

Problem:

What is the length of a tube that has a fundamental frequency of 176 Hz and a first overtone of 352 Hz if the speed of sound is 343 m/s?

Exercise:

Problem:

(a) Find the length of an organ pipe closed at one end that produces a fundamental frequency of 256 Hz when air temperature is $18.0^{\circ}\mathrm{C}$. (b) What is its fundamental frequency at $25.0^{\circ}\mathrm{C}$?

Solution:

(a) 0.334 m

(b) 259 Hz

Exercise:

Problem:

By what fraction will the frequencies produced by a wind instrument change when air temperature goes from $10.0^{\circ}C$ to $30.0^{\circ}C$? That is, find the ratio of the frequencies at those temperatures.

Exercise:

Problem:

The ear canal resonates like a tube closed at one end. (See [link].) If ear canals range in length from 1.80 to 2.60 cm in an average population, what is the range of fundamental resonant frequencies? Take air temperature to be $37.0^{\circ}C$, which is the same as body temperature. How does this result correlate with the intensity versus frequency graph ([link] of the human ear?

Solution:

3.39 to 4.90 kHz

Exercise:

Problem:

Calculate the first overtone in an ear canal, which resonates like a 2.40-cm-long tube closed at one end, by taking air temperature to be $37.0^{\circ}C$. Is the ear particularly sensitive to such a frequency? (The resonances of the ear canal are complicated by its nonuniform shape, which we shall ignore.)

Exercise:

Problem:

A crude approximation of voice production is to consider the breathing passages and mouth to be a resonating tube closed at one end. (See [link].) (a) What is the fundamental frequency if the tube is 0.240-m long, by taking air temperature to be $37.0^{\circ}C$? (b) What would this frequency become if the person replaced the air with helium? Assume the same temperature dependence for helium as for air.

Solution:

(a) 367 Hz

(b) 1.07 kHz

Exercise:

Problem:

(a) Students in a physics lab are asked to find the length of an air column in a tube closed at one end that has a fundamental frequency of 256 Hz. They hold the tube vertically and fill it with water to the top, then lower the water while a 256-Hz tuning fork is rung and listen for the first resonance. What is the air temperature if the resonance occurs for a length of 0.336 m? (b) At what length will they observe the second resonance (first overtone)?

Exercise:

Problem:

What frequencies will a 1.80-m-long tube produce in the audible range at $20.0^{\circ}\mathrm{C}$ if: (a) The tube is closed at one end? (b) It is open at both ends?

Solution:

(a) $f_n = n(47.6\text{ Hz}),\ n = 1, 3, 5,..., 419$

(b) $f_n = n(95.3\text{ Hz}),\ n = 1, 2, 3,..., 210$

Glossary

antinode
: point of maximum displacement

node

point of zero displacement

fundamental
the lowest-frequency resonance

overtones
all resonant frequencies higher than the fundamental

harmonics
the term used to refer collectively to the fundamental and its overtones

Hearing

- Define hearing, pitch, loudness, timbre, note, tone, phon, ultrasound, and infrasound.
- Compare loudness to frequency and intensity of a sound.
- Identify structures of the inner ear and explain how they relate to sound perception.

Hearing allows this vocalist, his band, and his fans to enjoy music. (credit: West Point Public Affairs, Flickr)

The human ear has a tremendous range and sensitivity. It can give us a wealth of simple information—such as pitch, loudness, and direction. And from its input we can detect musical quality and nuances of voiced emotion. How is our hearing related to the physical qualities of sound, and how does the hearing mechanism work?

Hearing is the perception of sound. (Perception is commonly defined to be awareness through the senses, a typically circular definition of higher-level processes in living organisms.) Normal human hearing encompasses frequencies from 20 to 20,000 Hz, an impressive range. Sounds below 20 Hz are called **infrasound**, whereas those above 20,000 Hz are **ultrasound**. Neither is perceived by the ear, although infrasound can sometimes be felt as vibrations. When we do hear low-frequency vibrations, such as the

sounds of a diving board, we hear the individual vibrations only because there are higher-frequency sounds in each. Other animals have hearing ranges different from that of humans. Dogs can hear sounds as high as 30,000 Hz, whereas bats and dolphins can hear up to 100,000-Hz sounds. You may have noticed that dogs respond to the sound of a dog whistle which produces sound out of the range of human hearing. Elephants are known to respond to frequencies below 20 Hz.

The perception of frequency is called **pitch**. Most of us have excellent relative pitch, which means that we can tell whether one sound has a different frequency from another. Typically, we can discriminate between two sounds if their frequencies differ by 0.3% or more. For example, 500.0 and 501.5 Hz are noticeably different. Pitch perception is directly related to frequency and is not greatly affected by other physical quantities such as intensity. Musical **notes** are particular sounds that can be produced by most instruments and in Western music have particular names. Combinations of notes constitute music. Some people can identify musical notes, such as A-sharp, C, or E-flat, just by listening to them. This uncommon ability is called perfect pitch.

The ear is remarkably sensitive to low-intensity sounds. The lowest audible intensity or threshold is about $10^{-12}\ \mathrm{W/m^2}$ or 0 dB. Sounds as much as 10^{12} more intense can be briefly tolerated. Very few measuring devices are capable of observations over a range of a trillion. The perception of intensity is called **loudness**. At a given frequency, it is possible to discern differences of about 1 dB, and a change of 3 dB is easily noticed. But loudness is not related to intensity alone. Frequency has a major effect on how loud a sound seems. The ear has its maximum sensitivity to frequencies in the range of 2000 to 5000 Hz, so that sounds in this range are perceived as being louder than, say, those at 500 or 10,000 Hz, even when they all have the same intensity. Sounds near the high- and low-frequency extremes of the hearing range seem even less loud, because the ear is even less sensitive at those frequencies. [link] gives the dependence of certain human hearing perceptions on physical quantities.

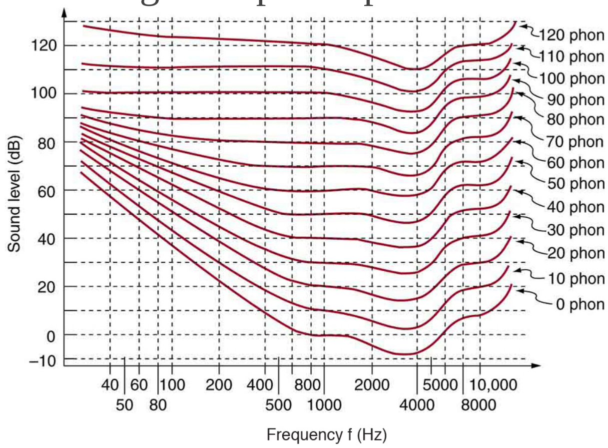
Perception	**Physical quantity**
Pitch	Frequency
Loudness	Intensity and Frequency
Timbre	Number and relative intensity of multiple frequencies. Subtle craftsmanship leads to non-linear effects and more detail.
Note	Basic unit of music with specific names, combined to generate tunes
Tone	Number and relative intensity of multiple frequencies.

Sound Perceptions

When a violin plays middle C, there is no mistaking it for a piano playing the same note. The reason is that each instrument produces a distinctive set of frequencies and intensities. We call our perception of these combinations of frequencies and intensities **tone** quality, or more commonly the **timbre** of the sound. It is more difficult to correlate timbre perception to physical quantities than it is for loudness or pitch perception. Timbre is more subjective. Terms such as dull, brilliant, warm, cold, pure, and rich are employed to describe the timbre of a sound. So the consideration of timbre takes us into the realm of perceptual psychology, where higher-level processes in the brain are dominant. This is true for other perceptions of sound, such as music and noise. We shall not delve further into them; rather, we will concentrate on the question of loudness perception.

A unit called a **phon** is used to express loudness numerically. Phons differ from decibels because the phon is a unit of loudness perception, whereas the decibel is a unit of physical intensity. [link] shows the relationship of loudness to intensity (or intensity level) and frequency for persons with normal hearing. The curved lines are equal-loudness curves. Each curve is

labeled with its loudness in phons. Any sound along a given curve will be perceived as equally loud by the average person. The curves were determined by having large numbers of people compare the loudness of sounds at different frequencies and sound intensity levels. At a frequency of 1000 Hz, phons are taken to be numerically equal to decibels. The following example helps illustrate how to use the graph:



The relationship of loudness in phons to intensity level (in decibels) and intensity (in watts per meter squared) for persons with normal hearing. The curved lines are equal-loudness curves—all sounds on a given curve are perceived as equally loud. Phons and decibels are defined to be the same at 1000 Hz.

Example:

Measuring Loudness: Loudness Versus Intensity Level and Frequency

(a) What is the loudness in phons of a 100-Hz sound that has an intensity level of 80 dB? (b) What is the intensity level in decibels of a 4000-Hz

sound having a loudness of 70 phons? (c) At what intensity level will an 8000-Hz sound have the same loudness as a 200-Hz sound at 60 dB?

Strategy for (a)

The graph in [link] should be referenced in order to solve this example. To find the loudness of a given sound, you must know its frequency and intensity level and locate that point on the square grid, then interpolate between loudness curves to get the loudness in phons.

Solution for (a)

(1) Identify knowns:

- The square grid of the graph relating phons and decibels is a plot of intensity level versus frequency—both physical quantities.
- 100 Hz at 80 dB lies halfway between the curves marked 70 and 80 phons.

(2) Find the loudness: 75 phons.

Strategy for (b)

The graph in [link] should be referenced in order to solve this example. To find the intensity level of a sound, you must have its frequency and loudness. Once that point is located, the intensity level can be determined from the vertical axis.

Solution for (b)

(1) Identify knowns:

- Values are given to be 4000 Hz at 70 phons.

(2) Follow the 70-phon curve until it reaches 4000 Hz. At that point, it is below the 70 dB line at about 67 dB.

(3) Find the intensity level:

67 dB

Strategy for (c)

The graph in [link] should be referenced in order to solve this example.

Solution for (c)

(1) Locate the point for a 200 Hz and 60 dB sound.

(2) Find the loudness: This point lies just slightly above the 50-phon curve, and so its loudness is 51 phons.

(3) Look for the 51-phon level is at 8000 Hz: 63 dB.

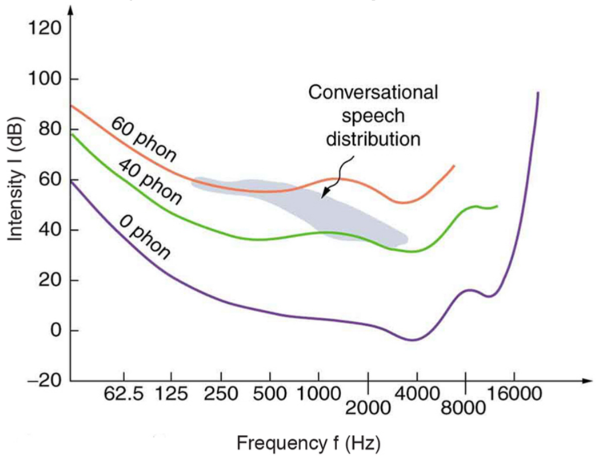
Discussion

These answers, like all information extracted from [link], have uncertainties of several phons or several decibels, partly due to difficulties in interpolation, but mostly related to uncertainties in the equal-loudness curves.

Further examination of the graph in [link] reveals some interesting facts about human hearing. First, sounds below the 0-phon curve are not perceived by most people. So, for example, a 60 Hz sound at 40 dB is inaudible. The 0-phon curve represents the threshold of normal hearing. We can hear some sounds at intensity levels below 0 dB. For example, a 3-dB, 5000-Hz sound is audible, because it lies above the 0-phon curve. The loudness curves all have dips in them between about 2000 and 5000 Hz. These dips mean the ear is most sensitive to frequencies in that range. For example, a 15-dB sound at 4000 Hz has a loudness of 20 phons, the same as a 20-dB sound at 1000 Hz. The curves rise at both extremes of the frequency range, indicating that a greater-intensity level sound is needed at those frequencies to be perceived to be as loud as at middle frequencies. For example, a sound at 10,000 Hz must have an intensity level of 30 dB to seem as loud as a 20 dB sound at 1000 Hz. Sounds above 120 phons are painful as well as damaging.

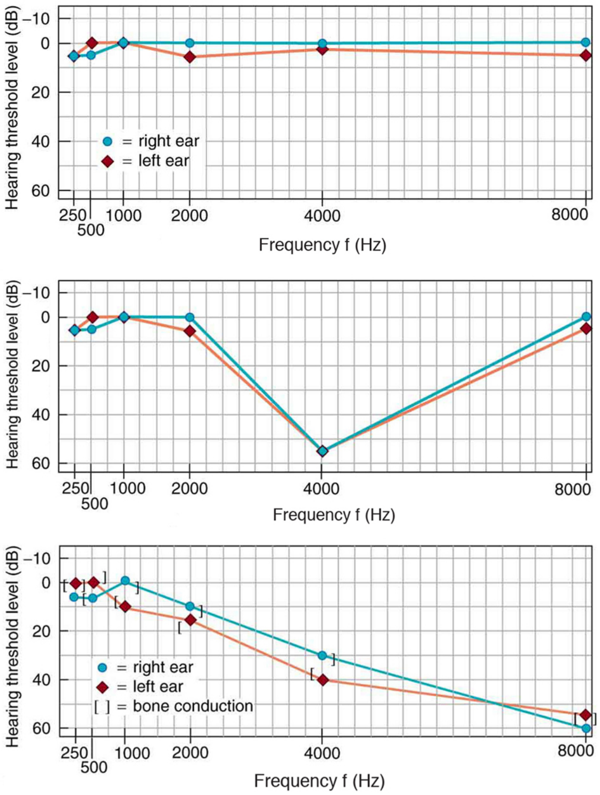
We do not often utilize our full range of hearing. This is particularly true for frequencies above 8000 Hz, which are rare in the environment and are unnecessary for understanding conversation or appreciating music. In fact, people who have lost the ability to hear such high frequencies are usually unaware of their loss until tested. The shaded region in [link] is the frequency and intensity region where most conversational sounds fall. The curved lines indicate what effect hearing losses of 40 and 60 phons will have. A 40-phon hearing loss at all frequencies still allows a person to understand conversation, although it will seem very quiet. A person with a 60-phon loss at all frequencies will hear only the lowest frequencies and will not be able to understand speech unless it is much louder than normal. Even so, speech may seem indistinct, because higher frequencies are not as well perceived. The conversational speech region also has a gender component, in that female voices are usually characterized by higher

frequencies. So the person with a 60-phon hearing impediment might have difficulty understanding the normal conversation of a woman.



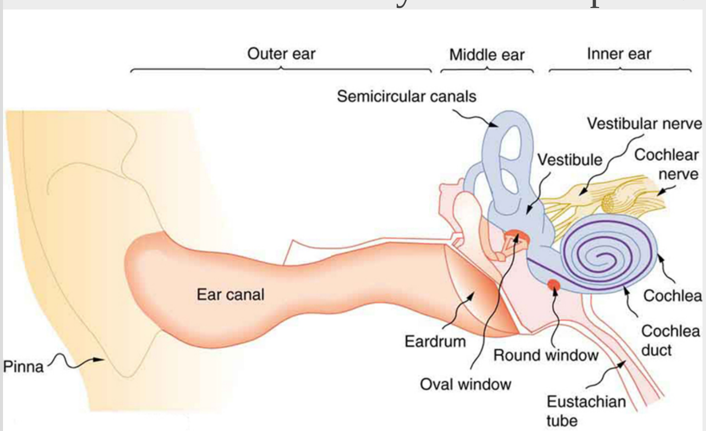
The shaded region represents frequencies and intensity levels found in normal conversational speech. The 0-phon line represents the normal hearing threshold, while those at 40 and 60 represent thresholds for people with 40- and 60-phon hearing losses, respectively.

Hearing tests are performed over a range of frequencies, usually from 250 to 8000 Hz, and can be displayed graphically in an audiogram like that in [link]. The hearing threshold is measured in dB *relative to the normal threshold*, so that normal hearing registers as 0 dB at all frequencies. Hearing loss caused by noise typically shows a dip near the 4000 Hz frequency, irrespective of the frequency that caused the loss and often affects both ears. The most common form of hearing loss comes with age and is called *presbycusis*—literally elder ear. Such loss is increasingly severe at higher frequencies, and interferes with music appreciation and speech recognition.



Audiograms showing the threshold in intensity level versus frequency for three different individuals. Intensity level is measured relative to the normal threshold. The top left graph is that of a person with normal hearing. The graph to its right has a dip at 4000 Hz and is that of a child who suffered hearing loss due to a cap gun. The third graph is typical of presbycusis, the progressive loss of higher frequency hearing with age. Tests performed by bone conduction (brackets) can distinguish nerve damage from middle ear damage.

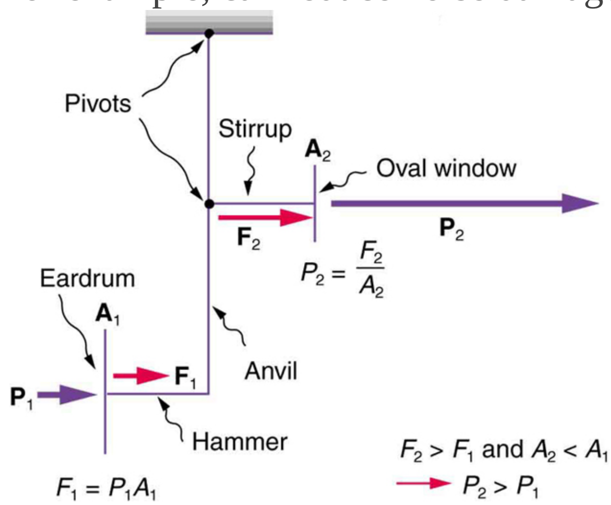
Note:
The Hearing Mechanism
The hearing mechanism involves some interesting physics. The sound wave that impinges upon our ear is a pressure wave. The ear is a transducer that converts sound waves into electrical nerve impulses in a manner much more sophisticated than, but analogous to, a microphone. [link] shows the gross anatomy of the ear with its division into three parts: the outer ear or ear canal; the middle ear, which runs from the eardrum to the cochlea; and the inner ear, which is the cochlea itself. The body part normally referred to as the ear is technically called the pinna.



The illustration shows the gross anatomy of the human ear.

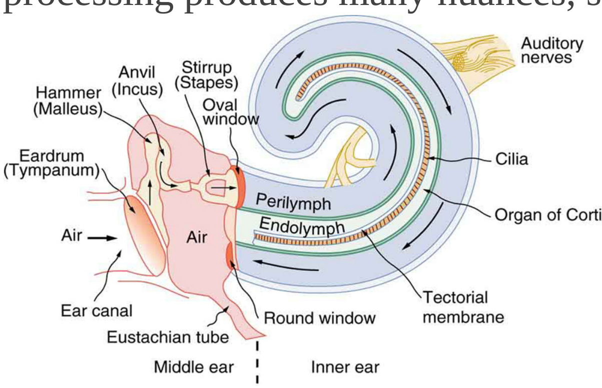
The outer ear, or ear canal, carries sound to the recessed protected eardrum. The air column in the ear canal resonates and is partially responsible for the sensitivity of the ear to sounds in the 2000 to 5000 Hz range. The middle ear converts sound into mechanical vibrations and applies these vibrations to the cochlea. The lever system of the middle ear takes the force exerted on the eardrum by sound pressure variations, amplifies it and transmits it to the

inner ear via the oval window, creating pressure waves in the cochlea approximately 40 times greater than those impinging on the eardrum. (See [link].) Two muscles in the middle ear (not shown) protect the inner ear from very intense sounds. They react to intense sound in a few milliseconds and reduce the force transmitted to the cochlea. This protective reaction can also be triggered by your own voice, so that humming while shooting a gun, for example, can reduce noise damage.



This schematic shows the middle ear’s system for converting sound pressure into force, increasing that force through a lever system, and applying the increased force to a small area of the cochlea, thereby creating a pressure about 40 times that in the original sound wave. A protective muscle reaction to intense sounds greatly reduces the mechanical advantage of the lever system.

[link] shows the middle and inner ear in greater detail. Pressure waves moving through the cochlea cause the tectorial membrane to vibrate, rubbing cilia (called hair cells), which stimulate nerves that send electrical signals to the brain. The membrane resonates at different positions for different frequencies, with high frequencies stimulating nerves at the near end and low frequencies at the far end. The complete operation of the cochlea is still not understood, but several mechanisms for sending information to the brain are known to be involved. For sounds below about 1000 Hz, the nerves send signals at the same frequency as the sound. For frequencies greater than about 1000 Hz, the nerves signal frequency by position. There is a structure to the cilia, and there are connections between nerve cells that perform signal processing before information is sent to the brain. Intensity information is partly indicated by the number of nerve signals and by volleys of signals. The brain processes the cochlear nerve signals to provide additional information such as source direction (based on time and intensity comparisons of sounds from both ears). Higher-level processing produces many nuances, such as music appreciation.



The inner ear, or cochlea, is a coiled tube about 3 mm in diameter and 3 cm in length if uncoiled. When the oval window is forced inward, as shown, a pressure wave travels through the perilymph in the direction of the arrows, stimulating nerves at the base of cilia in the organ of Corti.

Hearing losses can occur because of problems in the middle or inner ear. Conductive losses in the middle ear can be partially overcome by sending sound vibrations to the cochlea through the skull. Hearing aids for this purpose usually press against the bone behind the ear, rather than simply amplifying the sound sent into the ear canal as many hearing aids do. Damage to the nerves in the cochlea is not repairable, but amplification can partially compensate. There is a risk that amplification will produce further damage. Another common failure in the cochlea is damage or loss of the cilia but with nerves remaining functional. Cochlear implants that stimulate the nerves directly are now available and widely accepted. Over 100,000 implants are in use, in about equal numbers of adults and children.

The cochlear implant was pioneered in Melbourne, Australia, by Graeme Clark in the 1970s for his deaf father. The implant consists of three external components and two internal components. The external components are a microphone for picking up sound and converting it into an electrical signal, a speech processor to select certain frequencies and a transmitter to transfer the signal to the internal components through electromagnetic induction. The internal components consist of a receiver/transmitter secured in the bone beneath the skin, which converts the signals into electric impulses and sends them through an internal cable to the cochlea and an array of about 24 electrodes wound through the cochlea. These electrodes in turn send the impulses directly into the brain. The electrodes basically emulate the cilia.

Exercise:

Check Your Understanding

Problem:

Are ultrasound and infrasound imperceptible to all hearing organisms? Explain your answer.

Solution:

No, the range of perceptible sound is based in the range of human hearing. Many other organisms perceive either infrasound or ultrasound.

Section Summary

- The range of audible frequencies is 20 to 20,000 Hz.
- Those sounds above 20,000 Hz are ultrasound, whereas those below 20 Hz are infrasound.
- The perception of frequency is pitch.
- The perception of intensity is loudness.
- Loudness has units of phons.

Conceptual Questions

Exercise:

Problem:

Why can a hearing test show that your threshold of hearing is 0 dB at 250 Hz, when [link] implies that no one can hear such a frequency at less than 20 dB?

Problems & Exercises

Exercise:

Problem:

The factor of 10^{-12} in the range of intensities to which the ear can respond, from threshold to that causing damage after brief exposure, is truly remarkable. If you could measure distances over the same range with a single instrument and the smallest distance you could measure was 1 mm, what would the largest be?

Solution:
Equation:

$$1 \times 10^{6}\ \text{km}$$

Exercise:

Problem:

The frequencies to which the ear responds vary by a factor of 10^3. Suppose the speedometer on your car measured speeds differing by the same factor of 10^3, and the greatest speed it reads is 90.0 mi/h. What would be the slowest nonzero speed it could read?

Exercise:

Problem:

What are the closest frequencies to 500 Hz that an average person can clearly distinguish as being different in frequency from 500 Hz? The sounds are not present simultaneously.

Solution:

498.5 or 501.5 Hz

Exercise:

Problem:

Can the average person tell that a 2002-Hz sound has a different frequency than a 1999-Hz sound without playing them simultaneously?

Exercise:

Problem:

If your radio is producing an average sound intensity level of 85 dB, what is the next lowest sound intensity level that is clearly less intense?

Solution:

82 dB

Exercise:

Problem:

Can you tell that your roommate turned up the sound on the TV if its average sound intensity level goes from 70 to 73 dB?

Exercise:

Problem:

Based on the graph in [link], what is the threshold of hearing in decibels for frequencies of 60, 400, 1000, 4000, and 15,000 Hz? Note that many AC electrical appliances produce 60 Hz, music is commonly 400 Hz, a reference frequency is 1000 Hz, your maximum sensitivity is near 4000 Hz, and many older TVs produce a 15,750 Hz whine.

Solution:

approximately 48, 9, 0, –7, and 20 dB, respectively

Exercise:

Problem:

What sound intensity levels must sounds of frequencies 60, 3000, and 8000 Hz have in order to have the same loudness as a 40-dB sound of frequency 1000 Hz (that is, to have a loudness of 40 phons)?

Exercise:

Problem:

What is the approximate sound intensity level in decibels of a 600-Hz tone if it has a loudness of 20 phons? If it has a loudness of 70 phons?

Solution:

(a) 23 dB

(b) 70 dB

Exercise:

Problem:

(a) What are the loudnesses in phons of sounds having frequencies of 200, 1000, 5000, and 10,000 Hz, if they are all at the same 60.0-dB sound intensity level? (b) If they are all at 110 dB? (c) If they are all at 20.0 dB?

Exercise:

Problem:

Suppose a person has a 50-dB hearing loss at all frequencies. By how many factors of 10 will low-intensity sounds need to be amplified to seem normal to this person? Note that smaller amplification is appropriate for more intense sounds to avoid further hearing damage.

Solution:

Five factors of 10

Exercise:

Problem:

If a woman needs an amplification of 5.0×10^{12} times the threshold intensity to enable her to hear at all frequencies, what is her overall hearing loss in dB? Note that smaller amplification is appropriate for more intense sounds to avoid further damage to her hearing from levels above 90 dB.

Exercise:

Problem:

(a) What is the intensity in watts per meter squared of a just barely audible 200-Hz sound? (b) What is the intensity in watts per meter squared of a barely audible 4000-Hz sound?

Solution:

(a) $2 \times 10^{-10}\ \mathrm{W/m^2}$

(b) $2 \times 10^{-13}\ \mathrm{W/m^2}$

Exercise:

Problem:

(a) Find the intensity in watts per meter squared of a 60.0-Hz sound having a loudness of 60 phons. (b) Find the intensity in watts per meter squared of a 10,000-Hz sound having a loudness of 60 phons.

Exercise:

Problem:

A person has a hearing threshold 10 dB above normal at 100 Hz and 50 dB above normal at 4000 Hz. How much more intense must a 100-Hz tone be than a 4000-Hz tone if they are both barely audible to this person?

Solution:

2.5

Exercise:

Problem:

A child has a hearing loss of 60 dB near 5000 Hz, due to noise exposure, and normal hearing elsewhere. How much more intense is a 5000-Hz tone than a 400-Hz tone if they are both barely audible to the child?

Exercise:

Problem:

What is the ratio of intensities of two sounds of identical frequency if the first is just barely discernible as louder to a person than the second?

Solution:

1.26

Glossary

loudness
: the perception of sound intensity

timbre
: number and relative intensity of multiple sound frequencies

note
: basic unit of music with specific names, combined to generate tunes

tone
: number and relative intensity of multiple sound frequencies

phon
: the numerical unit of loudness

ultrasound
: sounds above 20,000 Hz

infrasound
: sounds below 20 Hz

Ultrasound

- Define acoustic impedance and intensity reflection coefficient.
- Describe medical and other uses of ultrasound technology.
- Calculate acoustic impedance using density values and the speed of ultrasound.
- Calculate the velocity of a moving object using Doppler-shifted ultrasound.

Ultrasound is used in medicine to painlessly and noninvasively monitor patient health and diagnose a wide range of disorders. (credit: abbybatchelder, Flickr)

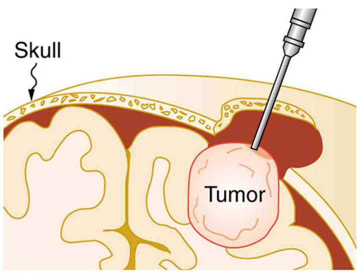
Any sound with a frequency above 20,000 Hz (or 20 kHz)—that is, above the highest audible frequency—is defined to be ultrasound. In practice, it is possible to create ultrasound frequencies up to more than a gigahertz. (Higher frequencies are difficult to create; furthermore, they propagate poorly because they are very strongly absorbed.) Ultrasound has a tremendous number of applications, which range from burglar alarms to use in cleaning delicate objects to the guidance systems of bats. We begin our discussion of ultrasound with some of its applications in medicine, in which it is used extensively both for diagnosis and for therapy.

Note:
Characteristics of Ultrasound
The characteristics of ultrasound, such as frequency and intensity, are wave properties common to all types of waves. Ultrasound also has a wavelength that limits the fineness of detail it can detect. This characteristic is true of all waves. We can never observe details significantly smaller than the wavelength of our probe; for example,

we will never see individual atoms with visible light, because the atoms are so small compared with the wavelength of light.

Ultrasound in Medical Therapy

Ultrasound, like any wave, carries energy that can be absorbed by the medium carrying it, producing effects that vary with intensity. When focused to intensities of 10^3 to $10^5\ \mathrm{W/m^2}$, ultrasound can be used to shatter gallstones or pulverize cancerous tissue in surgical procedures. (See [link].) Intensities this great can damage individual cells, variously causing their protoplasm to stream inside them, altering their permeability, or rupturing their walls through *cavitation*. Cavitation is the creation of vapor cavities in a fluid—the longitudinal vibrations in ultrasound alternatively compress and expand the medium, and at sufficient amplitudes the expansion separates molecules. Most cavitation damage is done when the cavities collapse, producing even greater shock pressures.



The tip of this small probe oscillates at 23 kHz with such a large amplitude that it pulverizes tissue on contact. The debris is then aspirated. The speed of the tip may exceed the speed of sound in tissue, thus creating shock waves and cavitation, rather than a smooth

simple harmonic
oscillator–type
wave.

Most of the energy carried by high-intensity ultrasound in tissue is converted to thermal energy. In fact, intensities of 10^3 to $10^4\ \mathrm{W/m^2}$ are commonly used for deep-heat treatments called ultrasound diathermy. Frequencies of 0.8 to 1 MHz are typical. In both athletics and physical therapy, ultrasound diathermy is most often applied to injured or overworked muscles to relieve pain and improve flexibility. Skill is needed by the therapist to avoid “bone burns” and other tissue damage caused by overheating and cavitation, sometimes made worse by reflection and focusing of the ultrasound by joint and bone tissue.

In some instances, you may encounter a different decibel scale, called the sound *pressure* level, when ultrasound travels in water or in human and other biological tissues. We shall not use the scale here, but it is notable that numbers for sound pressure levels range 60 to 70 dB higher than you would quote for β, the sound intensity level used in this text. Should you encounter a sound pressure level of 220 decibels, then, it is not an astronomically high intensity, but equivalent to about 155 dB—high enough to destroy tissue, but not as unreasonably high as it might seem at first.

Ultrasound in Medical Diagnostics

When used for imaging, ultrasonic waves are emitted from a transducer, a crystal exhibiting the piezoelectric effect (the expansion and contraction of a substance when a voltage is applied across it, causing a vibration of the crystal). These high-frequency vibrations are transmitted into any tissue in contact with the transducer. Similarly, if a pressure is applied to the crystal (in the form of a wave reflected off tissue layers), a voltage is produced which can be recorded. The crystal therefore acts as both a transmitter and a receiver of sound. Ultrasound is also partially absorbed by tissue on its path, both on its journey away from the transducer and on its return journey. From the time between when the original signal is sent and when the reflections from various boundaries between media are received, (as well as a measure of the intensity loss of the signal), the nature and position of each boundary between tissues and organs may be deduced.

Reflections at boundaries between two different media occur because of differences in a characteristic known as the **acoustic impedance** Z of each substance. Impedance is defined as

Equation:

$$Z = \rho v,$$

where ρ is the density of the medium (in kg/m^3) and v is the speed of sound through the medium (in m/s). The units for Z are therefore $kg/(m^2 \cdot s)$.

[link] shows the density and speed of sound through various media (including various soft tissues) and the associated acoustic impedances. Note that the acoustic impedances for soft tissue do not vary much but that there is a big difference between the acoustic impedance of soft tissue and air and also between soft tissue and bone.

Medium	**Density (kg/m^3)**	**Speed of Ultrasound (m/s)**	**Acoustic Impedance** $\left(kg/\left(m^2 \cdot s\right)\right)$
Air	1.3	330	429
Water	1000	1500	1.5×10^6
Blood	1060	1570	1.66×10^6
Fat	925	1450	1.34×10^6
Muscle (average)	1075	1590	1.70×10^6
Bone (varies)	1400–1900	4080	5.7×10^6 to 7.8×10^6
Barium titanate (transducer material)	5600	5500	30.8×10^6

The Ultrasound Properties of Various Media, Including Soft Tissue Found in the Body

At the boundary between media of different acoustic impedances, some of the wave energy is reflected and some is transmitted. The greater the *difference* in acoustic impedance between the two media, the greater the reflection and the smaller the transmission.

The **intensity reflection coefficient** a is defined as the ratio of the intensity of the reflected wave relative to the incident (transmitted) wave. This statement can be written mathematically as

Equation:

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2},$$

where Z_1 and Z_2 are the acoustic impedances of the two media making up the boundary. A reflection coefficient of zero (corresponding to total transmission and no reflection) occurs when the acoustic impedances of the two media are the same. An impedance "match" (no reflection) provides an efficient coupling of sound energy from one medium to another. The image formed in an ultrasound is made by tracking reflections (as shown in [link]) and mapping the intensity of the reflected sound waves in a two-dimensional plane.

Example:

Calculate Acoustic Impedance and Intensity Reflection Coefficient: Ultrasound and Fat Tissue

(a) Using the values for density and the speed of ultrasound given in [link], show that the acoustic impedance of fat tissue is indeed $1.34 \times 10^6 \text{ kg/(m}^2\text{·s)}$.

(b) Calculate the intensity reflection coefficient of ultrasound when going from fat to muscle tissue.

Strategy for (a)

The acoustic impedance can be calculated using $Z = \rho v$ and the values for ρ and v found in [link].

Solution for (a)

(1) Substitute known values from [link] into $Z = \rho v$.

Equation:

$$Z = \rho v = \left(925 \text{ kg/m}^3\right)(1450 \text{ m/s})$$

(2) Calculate to find the acoustic impedance of fat tissue.

Equation:

$$1.34 \times 10^6\ \text{kg/(m}^2\text{·s)}$$

This value is the same as the value given for the acoustic impedance of fat tissue.

Strategy for (b)

The intensity reflection coefficient for any boundary between two media is given by $a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$, and the acoustic impedance of muscle is given in [link].

Solution for (b)

Substitute known values into $a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$ to find the intensity reflection coefficient:

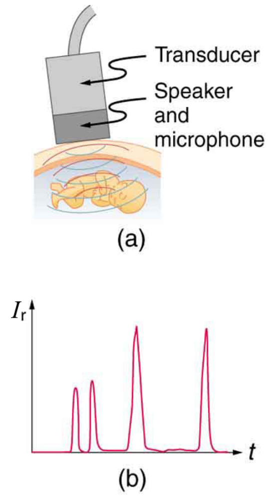
Equation:

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2} = \frac{\left(1.34 \times 10^6\ \text{kg/(m}^2\cdot\text{s)} - 1.70 \times 10^6\ \text{kg/(m}^2\cdot\text{s)}\right)^2}{\left(1.70 \times 10^6\ \text{kg/(m}^2\cdot\text{s)} + 1.34 \times 10^6\ \text{kg/(m}^2\cdot\text{s)}\right)^2} = 0.014$$

Discussion

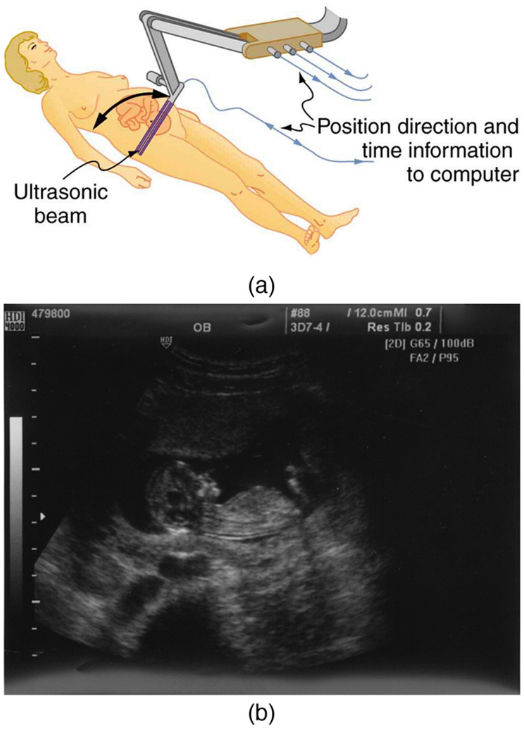
This result means that only 1.4% of the incident intensity is reflected, with the remaining being transmitted.

The applications of ultrasound in medical diagnostics have produced untold benefits with no known risks. Diagnostic intensities are too low (about $10^{-2}\ \text{W/m}^2$) to cause thermal damage. More significantly, ultrasound has been in use for several decades and detailed follow-up studies do not show evidence of ill effects, quite unlike the case for x-rays.



(a) An ultrasound speaker doubles as a microphone. Brief bleeps are broadcast, and echoes are recorded from various depths. (b) Graph of echo intensity versus time. The time for echoes to return is directly proportional to the distance of the reflector, yielding this information noninvasively.

The most common ultrasound applications produce an image like that shown in [link]. The speaker-microphone broadcasts a directional beam, sweeping the beam across the area of interest. This is accomplished by having multiple ultrasound sources in the probe's head, which are phased to interfere constructively in a given, adjustable direction. Echoes are measured as a function of position as well as depth. A computer constructs an image that reveals the shape and density of internal structures.



(a) An ultrasonic image is produced by sweeping the ultrasonic beam across the area of interest, in this case the woman's abdomen. Data are recorded and analyzed in a computer, providing a two-dimensional image. (b) Ultrasound image of 12-week-old fetus. (credit: Margaret W. Carruthers, Flickr)

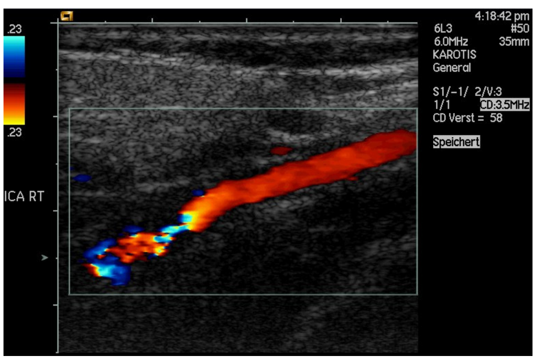
How much detail can ultrasound reveal? The image in [link] is typical of low-cost systems, but that in [link] shows the remarkable detail possible with more advanced systems, including 3D imaging. Ultrasound today is commonly used in prenatal care. Such imaging can be used to see if the fetus is developing at a normal rate, and help in the determination of serious problems early in the pregnancy. Ultrasound is also in wide use to image the chambers of the heart and the flow of blood within the beating heart, using the Doppler effect (echocardiology).

Whenever a wave is used as a probe, it is very difficult to detect details smaller than its wavelength λ. Indeed, current technology cannot do quite this well. Abdominal scans may use a 7-MHz frequency, and the speed of sound in tissue is about 1540 m/s —so the wavelength limit to detail would be $\lambda = \frac{v_{\text{w}}}{f} = \frac{1540 \text{ m/s}}{7 \times 10^6 \text{ Hz}} = 0.22 \text{ mm}$. In practice, 1-mm detail is attainable, which is sufficient for many purposes. Higher-frequency ultrasound would allow greater detail, but it does not penetrate as well as lower frequencies do. The accepted rule of thumb is that you can effectively scan to a depth of about 500λ into tissue. For 7 MHz, this penetration limit is $500 \times 0.22 \text{ mm}$, which is 0.11 m. Higher frequencies may be employed in smaller organs, such as the eye, but are not practical for looking deep into the body.

A 3D ultrasound image of a fetus. As well as for the detection of any abnormalities, such scans have also been shown to be useful for strengthening the emotional bonding between parents and their unborn child. (credit: Jennie Cu, Wikimedia Commons)

In addition to shape information, ultrasonic scans can produce density information superior to that found in X-rays, because the intensity of a reflected sound is related to changes in density. Sound is most strongly reflected at places where density changes are greatest.

Another major use of ultrasound in medical diagnostics is to detect motion and determine velocity through the Doppler shift of an echo, known as **Doppler-shifted ultrasound**. This technique is used to monitor fetal heartbeat, measure blood velocity, and detect occlusions in blood vessels, for example. (See [link].) The magnitude of the Doppler shift in an echo is directly proportional to the velocity of whatever reflects the sound. Because an echo is involved, there is actually a double shift. The first occurs because the reflector (say a fetal heart) is a moving observer and receives a Doppler-shifted frequency. The reflector then acts as a moving source, producing a second Doppler shift.



This Doppler-shifted ultrasonic image of a partially occluded artery uses color to indicate velocity. The highest velocities are in red, while the lowest are blue. The blood must move faster through the constriction to carry the same flow. (credit: Arning C, Grzyska U, Wikimedia Commons)

A clever technique is used to measure the Doppler shift in an echo. The frequency of the echoed sound is superimposed on the broadcast frequency, producing beats. The beat frequency is $F_{\text{B}} = | f_1 - f_2 |$, and so it is directly proportional to the Doppler shift ($f_1 - f_2$) and hence, the reflector's velocity. The advantage in this technique is that the Doppler shift is small (because the reflector's velocity is small), so that great accuracy would be needed to measure the shift directly. But measuring the beat frequency is easy, and it is not affected if the broadcast frequency varies somewhat. Furthermore, the beat frequency is in the audible range and can be amplified for audio feedback to the medical observer.

Note:

Uses for Doppler-Shifted Radar

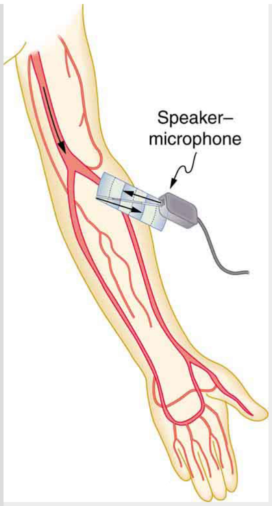
Doppler-shifted radar echoes are used to measure wind velocities in storms as well as aircraft and automobile speeds. The principle is the same as for Doppler-shifted ultrasound. There is evidence that bats and dolphins may also sense the velocity of an object (such as prey) reflecting their ultrasound signals by observing its Doppler shift.

Example:

Calculate Velocity of Blood: Doppler-Shifted Ultrasound

Ultrasound that has a frequency of 2.50 MHz is sent toward blood in an artery that is moving toward the source at 20.0 cm/s, as illustrated in [link]. Use the speed of sound in human tissue as 1540 m/s. (Assume that the frequency of 2.50 MHz is accurate to seven significant figures.)

a. What frequency does the blood receive?
b. What frequency returns to the source?
c. What beat frequency is produced if the source and returning frequencies are mixed?



Ultrasound is partly reflected by blood cells and plasma back toward the speaker-microphone. Because the cells are moving, two Doppler shifts are produced—one for blood as a moving observer, and the other for the reflected sound coming from a moving source. The magnitude of the shift is directly proportional to blood velocity.

Strategy

The first two questions can be answered using $f_{obs} = f_s\left(\frac{v_w}{v_w \pm v_s}\right)$ and $f_{obs} = f_s\left(\frac{v_w \pm v_{obs}}{v_w}\right)$ for the Doppler shift. The last question asks for beat frequency, which is the difference between the original and returning frequencies.

Solution for (a)

(1) Identify knowns:

- The blood is a moving observer, and so the frequency it receives is given by
  Equation:
  $$f_{obs} = f_s\left(\frac{v_w \pm v_{obs}}{v_w}\right).$$
- v_b is the blood velocity (v_{obs} here) and the plus sign is chosen because the motion is toward the source.

(2) Enter the given values into the equation.

Equation:

$$f_{obs} = (2{,}500{,}000\ \text{Hz})\left(\frac{1540\ \text{m/s} + 0.2\ \text{m/s}}{1540\ \text{m/s}}\right)$$

(3) Calculate to find the frequency: 2,500,325 Hz.

Solution for (b)

(1) Identify knowns:

- The blood acts as a moving source.
- The microphone acts as a stationary observer.
- The frequency leaving the blood is 2,500,325 Hz, but it is shifted upward as given by
  Equation:
  $$f_{obs} = f_s\left(\frac{v_w}{v_w - v_b}\right).$$
  f_{obs} is the frequency received by the speaker-microphone.
- The source velocity is v_b.
- The minus sign is used because the motion is toward the observer.

The minus sign is used because the motion is toward the observer.
(2) Enter the given values into the equation:
Equation:

$$f_{\text{obs}} = (2{,}500{,}325\ \text{Hz})\left(\frac{1540\ \text{m/s}}{1540\ \text{m/s} - 0.200\ \text{m/s}}\right)$$

(3) Calculate to find the frequency returning to the source: 2,500,649 Hz.
Solution for (c)
(1) Identify knowns:

- The beat frequency is simply the absolute value of the difference between f_{s} and f_{obs}, as stated in:
  Equation:

  $$f_{\text{B}} = \mid f_{\text{obs}} - f_{\text{s}} \mid.$$

(2) Substitute known values:
Equation:

$$\mid 2{,}500{,}649\ \text{Hz} - 2{,}500{,}000\ \text{Hz} \mid$$

(3) Calculate to find the beat frequency: 649 Hz.
Discussion
The Doppler shifts are quite small compared with the original frequency of 2.50 MHz. It is far easier to measure the beat frequency than it is to measure the echo frequency with an accuracy great enough to see shifts of a few hundred hertz out of a couple of megahertz. Furthermore, variations in the source frequency do not greatly affect the beat frequency, because both f_{s} and f_{obs}would increase or decrease. Those changes subtract out in $f_{\text{B}} = \mid f_{\text{obs}} - f_{\text{s}} \mid$.

Note:
Industrial and Other Applications of Ultrasound
Industrial, retail, and research applications of ultrasound are common. A few are discussed here. Ultrasonic cleaners have many uses. Jewelry, machined parts, and other objects that have odd shapes and crevices are immersed in a cleaning fluid that is agitated with ultrasound typically about 40 kHz in frequency. The intensity is great enough to cause cavitation, which is responsible for most of the cleansing action. Because cavitation-produced shock pressures are large and well transmitted in a fluid,

they reach into small crevices where even a low-surface-tension cleaning fluid might not penetrate.

Sonar is a familiar application of ultrasound. Sonar typically employs ultrasonic frequencies in the range from 30.0 to 100 kHz. Bats, dolphins, submarines, and even some birds use ultrasonic sonar. Echoes are analyzed to give distance and size information both for guidance and finding prey. In most sonar applications, the sound reflects quite well because the objects of interest have significantly different density than the medium in which they travel. When the Doppler shift is observed, velocity information can also be obtained. Submarine sonar can be used to obtain such information, and there is evidence that some bats also sense velocity from their echoes.

Similarly, there are a range of relatively inexpensive devices that measure distance by timing ultrasonic echoes. Many cameras, for example, use such information to focus automatically. Some doors open when their ultrasonic ranging devices detect a nearby object, and certain home security lights turn on when their ultrasonic rangers observe motion. Ultrasonic “measuring tapes” also exist to measure such things as room dimensions. Sinks in public restrooms are sometimes automated with ultrasound devices to turn faucets on and off when people wash their hands. These devices reduce the spread of germs and can conserve water.

Ultrasound is used for nondestructive testing in industry and by the military. Because ultrasound reflects well from any large change in density, it can reveal cracks and voids in solids, such as aircraft wings, that are too small to be seen with x-rays. For similar reasons, ultrasound is also good for measuring the thickness of coatings, particularly where there are several layers involved.

Basic research in solid state physics employs ultrasound. Its attenuation is related to a number of physical characteristics, making it a useful probe. Among these characteristics are structural changes such as those found in liquid crystals, the transition of a material to a superconducting phase, as well as density and other properties.

These examples of the uses of ultrasound are meant to whet the appetites of the curious, as well as to illustrate the underlying physics of ultrasound. There are many more applications, as you can easily discover for yourself.

Exercise:
Check Your Understanding

Problem:

Why is it possible to use ultrasound both to observe a fetus in the womb and also to destroy cancerous tumors in the body?

Solution:

Ultrasound can be used medically at different intensities. Lower intensities do not cause damage and are used for medical imaging. Higher intensities can pulverize and destroy targeted substances in the body, such as tumors.

Section Summary

- The acoustic impedance is defined as:
  Equation:
  $$Z = \rho v,$$
  ρ is the density of a medium through which the sound travels and v is the speed of sound through that medium.
- The intensity reflection coefficient a, a measure of the ratio of the intensity of the wave reflected off a boundary between two media relative to the intensity of the incident wave, is given by
  Equation:
  $$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}.$$
- The intensity reflection coefficient is a unitless quantity.

Conceptual Questions

Exercise:

Problem:

If audible sound follows a rule of thumb similar to that for ultrasound, in terms of its absorption, would you expect the high or low frequencies from your neighbor’s stereo to penetrate into your house? How does this expectation compare with your experience?

Exercise:

Problem:

Elephants and whales are known to use infrasound to communicate over very large distances. What are the advantages of infrasound for long distance communication?

Exercise:

Problem:

It is more difficult to obtain a high-resolution ultrasound image in the abdominal region of someone who is overweight than for someone who has a slight build. Explain why this statement is accurate.

Exercise:

Problem:

Suppose you read that 210-dB ultrasound is being used to pulverize cancerous tumors. You calculate the intensity in watts per centimeter squared and find it is unreasonably high (10^5 W/cm^2). What is a possible explanation?

Problems & Exercises

Unless otherwise indicated, for problems in this section, assume that the speed of sound through human tissues is 1540 m/s.

Exercise:

Problem:

What is the sound intensity level in decibels of ultrasound of intensity 10^5 W/m^2, used to pulverize tissue during surgery?

Solution:

170 dB

Exercise:

Problem:

Is 155-dB ultrasound in the range of intensities used for deep heating? Calculate the intensity of this ultrasound and compare this intensity with values quoted in the text.

Exercise:

Problem:

Find the sound intensity level in decibels of $2.00 \times 10^{-2} \text{ W/m}^2$ ultrasound used in medical diagnostics.

Solution:

103 dB

Exercise:

Problem:

The time delay between transmission and the arrival of the reflected wave of a signal using ultrasound traveling through a piece of fat tissue was 0.13 ms. At what depth did this reflection occur?

Exercise:

Problem:

In the clinical use of ultrasound, transducers are always coupled to the skin by a thin layer of gel or oil, replacing the air that would otherwise exist between the transducer and the skin. (a) Using the values of acoustic impedance given in [link] calculate the intensity reflection coefficient between transducer material and air. (b) Calculate the intensity reflection coefficient between transducer material and gel (assuming for this problem that its acoustic impedance is identical to that of water). (c) Based on the results of your calculations, explain why the gel is used.

Solution:

(a) 1.00

(b) 0.823

(c) Gel is used to facilitate the transmission of the ultrasound between the transducer and the patient’s body.

Exercise:

Problem:

(a) Calculate the minimum frequency of ultrasound that will allow you to see details as small as 0.250 mm in human tissue. (b) What is the effective depth to which this sound is effective as a diagnostic probe?

Exercise:

Problem:

(a) Find the size of the smallest detail observable in human tissue with 20.0-MHz ultrasound. (b) Is its effective penetration depth great enough to examine the entire eye (about 3.00 cm is needed)? (c) What is the wavelength of such ultrasound in $0^{\circ}\mathrm{C}$ air?

Solution:

(a) $77.0\ \mu\mathrm{m}$

(b) Effective penetration depth = 3.85 cm, which is enough to examine the eye.

(c) $16.6\ \mu\mathrm{m}$

Exercise:

Problem:

(a) Echo times are measured by diagnostic ultrasound scanners to determine distances to reflecting surfaces in a patient. What is the difference in echo times for tissues that are 3.50 and 3.60 cm beneath the surface? (This difference is the minimum resolving time for the scanner to see details as small as 0.100 cm, or 1.00 mm. Discrimination of smaller time differences is needed to see smaller details.) (b) Discuss whether the period T of this ultrasound must be smaller than the minimum time resolution. If so, what is the minimum frequency of the ultrasound and is that out of the normal range for diagnostic ultrasound?

Exercise:

Problem:

(a) How far apart are two layers of tissue that produce echoes having round-trip times (used to measure distances) that differ by $0.750\ \mu\mathrm{s}$? (b) What minimum frequency must the ultrasound have to see detail this small?

Solution:

(a) $5.78 \times 10^{-4}\ \mathrm{m}$

(b) $2.67 \times 10^{6}\ \mathrm{Hz}$

Exercise:

Problem:

(a) A bat uses ultrasound to find its way among trees. If this bat can detect echoes 1.00 ms apart, what minimum distance between objects can it detect? (b) Could this distance explain the difficulty that bats have finding an open door when they accidentally get into a house?

Exercise:

Problem:

A dolphin is able to tell in the dark that the ultrasound echoes received from two sharks come from two different objects only if the sharks are separated by 3.50 m, one being that much farther away than the other. (a) If the ultrasound has a frequency of 100 kHz, show this ability is not limited by its wavelength. (b) If this ability is due to the dolphin's ability to detect the arrival times of echoes, what is the minimum time difference the dolphin can perceive?

Solution:

(a) $v_{\text{w}} = 1540\text{ m/s} = f\lambda \Rightarrow \lambda = \frac{1540\text{ m/s}}{100\times10^{3}\text{ Hz}} = 0.0154\text{ m} < 3.50\text{ m}$. Because the wavelength is much shorter than the distance in question, the wavelength is not the limiting factor.

(b) 4.55 ms

Exercise:

Problem:

A diagnostic ultrasound echo is reflected from moving blood and returns with a frequency 500 Hz higher than its original 2.00 MHz. What is the velocity of the blood? (Assume that the frequency of 2.00 MHz is accurate to seven significant figures and 500 Hz is accurate to three significant figures.)

Exercise:

Problem:

Ultrasound reflected from an oncoming bloodstream that is moving at 30.0 cm/s is mixed with the original frequency of 2.50 MHz to produce beats. What is the beat frequency? (Assume that the frequency of 2.50 MHz is accurate to seven significant figures.)

Solution:

974 Hz

(Note: extra digits were retained in order to show the difference.)

Glossary

acoustic impedance
: property of medium that makes the propagation of sound waves more difficult

intensity reflection coefficient
: a measure of the ratio of the intensity of the wave reflected off a boundary between two media relative to the intensity of the incident wave

Doppler-shifted ultrasound
: a medical technique to detect motion and determine velocity through the Doppler shift of an echo

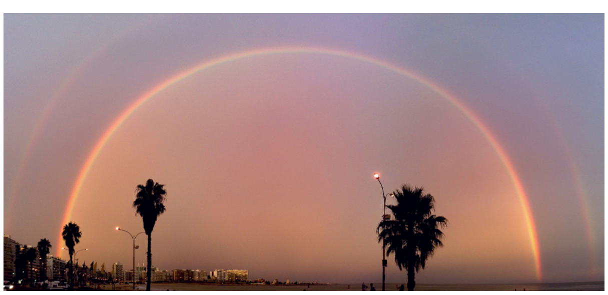
Introduction to Geometric Optics
class="introduction"

Geometric Optics

Light from this page or screen is formed into an image by the lens of your eye, much as the lens of the camera that made this photograph. Mirrors, like lenses, can also form images that in turn are captured by your eye.

Image seen as a result of reflection of light on a plane smooth surface. (credit: NASA Goddard Photo and Video, via Flickr)

Our lives are filled with light. Through vision, the most valued of our senses, light can evoke spiritual emotions, such as when we view a magnificent sunset or glimpse a rainbow breaking through the clouds. Light can also simply amuse us in a theater, or warn us to stop at an intersection. It has innumerable uses beyond vision. Light can carry telephone signals through glass fibers or cook a meal in a solar oven. Life itself could not exist without light's energy. From photosynthesis in plants to the sun warming a cold-blooded animal, its supply of energy is vital.



Double Rainbow over the bay

of Pocitos in Montevideo, Uruguay. (credit: Madrax, Wikimedia Commons)

We already know that visible light is the type of electromagnetic waves to which our eyes respond. That knowledge still leaves many questions regarding the nature of light and vision. What is color, and how do our eyes detect it? Why do diamonds sparkle? How does light travel? How do lenses and mirrors form images? These are but a few of the questions that are answered by the study of optics. Optics is the branch of physics that deals with the behavior of visible light and other electromagnetic waves. In particular, optics is concerned with the generation and propagation of light and its interaction with matter. What we have already learned about the generation of light in our study of heat transfer by radiation will be expanded upon in later topics, especially those on atomic physics. Now, we will concentrate on the propagation of light and its interaction with matter.

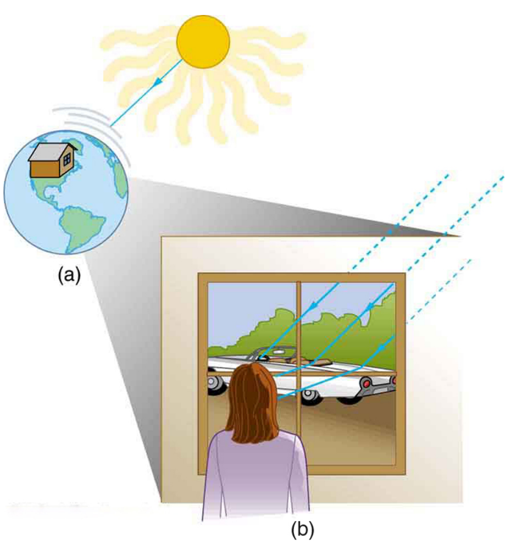
It is convenient to divide optics into two major parts based on the size of objects that light encounters. When light interacts with an object that is several times as large as the light's wavelength, its observable behavior is like that of a ray; it does not prominently display its wave characteristics. We call this part of optics "geometric optics." This chapter will concentrate on such situations. When light interacts with smaller objects, it has very prominent wave characteristics, such as constructive and destructive interference. Wave Optics will concentrate on such situations.

The Ray Aspect of Light

- List the ways by which light travels from a source to another location.

There are three ways in which light can travel from a source to another location. (See [link].) It can come directly from the source through empty space, such as from the Sun to Earth. Or light can travel through various media, such as air and glass, to the person. Light can also arrive after being reflected, such as by a mirror. In all of these cases, light is modeled as traveling in straight lines called rays. Light may change direction when it encounters objects (such as a mirror) or in passing from one material to another (such as in passing from air to glass), but it then continues in a straight line or as a ray. The word **ray** comes from mathematics and here means a straight line that originates at some point. It is acceptable to visualize light rays as laser rays (or even science fiction depictions of ray guns).

> **Note:**
> Ray
> The word "ray" comes from mathematics and here means a straight line that originates at some point.



Three methods for light to travel from a source to another location. (a) Light reaches the upper atmosphere of Earth traveling through empty space directly from the source. (b) Light can reach a person in one of two ways. It can travel through media like air and glass. It can also reflect from an object like a mirror. In the situations shown here, light interacts with objects large enough that it travels in straight lines, like a ray.

Experiments, as well as our own experiences, show that when light interacts with objects several times as large as its wavelength, it travels in straight lines and acts like a ray. Its wave characteristics are not pronounced in such situations. Since the wavelength of light is less than a micron (a thousandth of a millimeter), it acts like a ray in the many common situations in which it encounters objects larger than a micron. For example, when light encounters anything we can observe with unaided eyes, such as a mirror, it acts like a ray, with only subtle wave characteristics. We will concentrate on the ray characteristics in this chapter.

Since light moves in straight lines, changing directions when it interacts with materials, it is described by geometry and simple trigonometry. This part of optics, where the ray aspect of light dominates, is therefore called **geometric optics**. There are two laws that govern how light changes direction when it interacts with matter. These are the law of reflection, for

situations in which light bounces off matter, and the law of refraction, for situations in which light passes through matter.

Note:
Geometric Optics
The part of optics dealing with the ray aspect of light is called geometric optics.

Section Summary

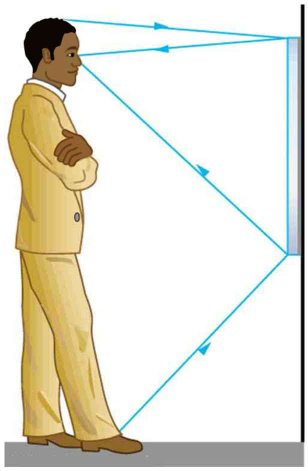
- A straight line that originates at some point is called a ray.
- The part of optics dealing with the ray aspect of light is called geometric optics.
- Light can travel in three ways from a source to another location: (1) directly from the source through empty space; (2) through various media; (3) after being reflected from a mirror.

Problems & Exercises

Exercise:

Problem:

Suppose a man stands in front of a mirror as shown in [link]. His eyes are 1.65 m above the floor, and the top of his head is 0.13 m higher. Find the height above the floor of the top and bottom of the smallest mirror in which he can see both the top of his head and his feet. How is this distance related to the man's height?



A full-length mirror is one in which you can see all of yourself. It need not be as big as you, and its size is independent of your distance from it.

Solution:

Top 1.715 m from floor, bottom 0.825 m from floor. Height of mirror is 0.890 m, or precisely one-half the height of the person.

Glossary

ray
: straight line that originates at some point

geometric optics

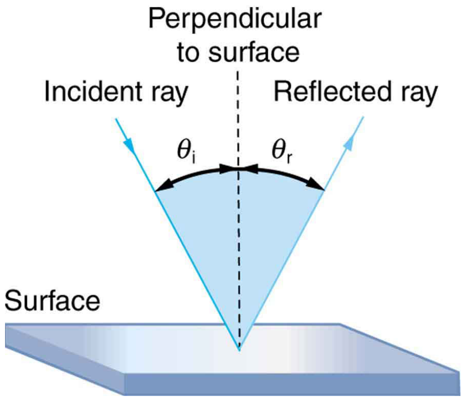
part of optics dealing with the ray aspect of light

The Law of Reflection

- Explain reflection of light from polished and rough surfaces.

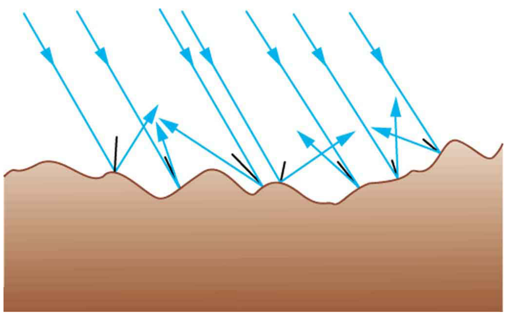
Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at this page, too, you are seeing light reflected from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

The law of reflection is illustrated in [link], which also shows how the angles are measured relative to the perpendicular to the surface at the point where the light ray strikes. We expect to see reflections from smooth surfaces, but [link] illustrates how a rough surface reflects light. Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. Diffused light is what allows us to see a sheet of paper from any angle, as illustrated in [link]. Many objects, such as people, clothing, leaves, and walls, have rough surfaces and can be seen from all sides. A mirror, on the other hand, has a smooth surface (compared with the wavelength of light) and reflects light at specific angles, as illustrated in [link]. When the moon reflects from a lake, as shown in [link], a combination of these effects takes place.

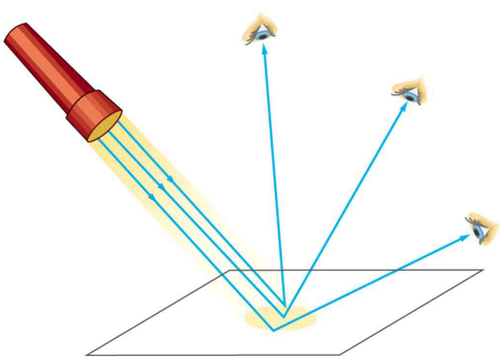


The law of reflection states that the angle of reflection equals the angle of incidence—$\theta_r = \theta_i$. The angles are measured relative to the perpendicular to

the surface at the point where the ray strikes the surface.

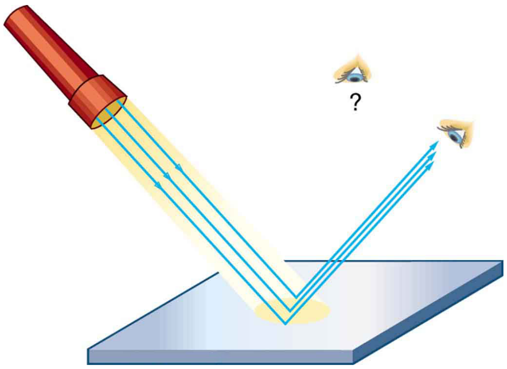


Light is diffused when it reflects from a rough surface. Here many parallel rays are incident, but they are reflected at many different angles since the surface is rough.



When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because

its surface is rough and diffuses the light.



A mirror illuminated by many parallel rays reflects them in only one direction, since its surface is very smooth. Only the observer at a particular angle will see the reflected light.

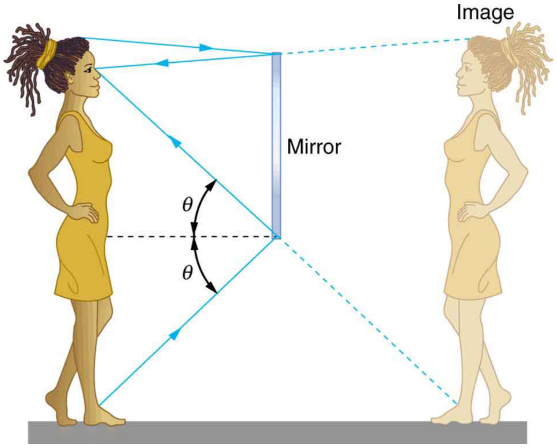
Moonlight is spread out when it is reflected by the lake, since the surface is shiny but uneven. (credit:

Diego Torres Silvestre,
Flickr)

The law of reflection is very simple: The angle of reflection equals the angle of incidence.

Note:
The Law of Reflection
The angle of reflection equals the angle of incidence.

When we see ourselves in a mirror, it appears that our image is actually behind the mirror. This is illustrated in [link]. We see the light coming from a direction determined by the law of reflection. The angles are such that our image is exactly the same distance behind the mirror as we stand away from the mirror. If the mirror is on the wall of a room, the images in it are all behind the mirror, which can make the room seem bigger. Although these mirror images make objects appear to be where they cannot be (like behind a solid wall), the images are not figments of our imagination. Mirror images can be photographed and videotaped by instruments and look just as they do with our eyes (optical instruments themselves). The precise manner in which images are formed by mirrors and lenses will be treated in later sections of this chapter.



Our image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be in the direction the rays are coming from when they enter the eyes.

Note:
Take-Home Experiment: Law of Reflection
Take a piece of paper and shine a flashlight at an angle at the paper, as shown in [link]. Now shine the flashlight at a mirror at an angle. Do your observations confirm the predictions in [link] and [link]? Shine the flashlight on various surfaces and determine whether the reflected light is diffuse or not. You can choose a shiny metallic lid of a pot or your skin. Using the mirror and flashlight, can you confirm the law of reflection? You will need to draw lines on a piece of paper showing the incident and reflected rays. (This part works even better if you use a laser pencil.)

Section Summary

- The angle of reflection equals the angle of incidence.
- A mirror has a smooth surface and reflects light at specific angles.
- Light is diffused when it reflects from a rough surface.
- Mirror images can be photographed and videotaped by instruments.

Conceptual Questions

Exercise:

Problem:

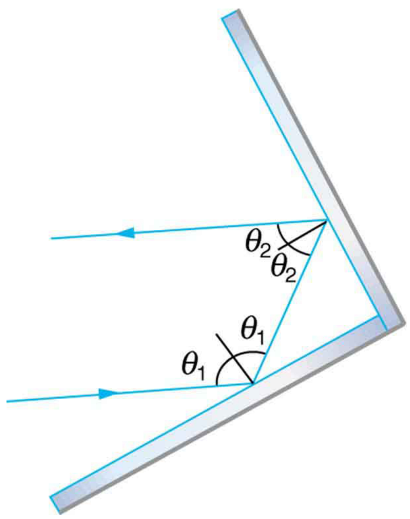
Using the law of reflection, explain how powder takes the shine off of a person’s nose. What is the name of the optical effect?

Problems & Exercises

Exercise:

Problem:

Show that when light reflects from two mirrors that meet each other at a right angle, the outgoing ray is parallel to the incoming ray, as illustrated in the following figure.



A corner reflector sends the reflected ray back in a direction parallel to the incident ray, independent of incoming direction.

Exercise:

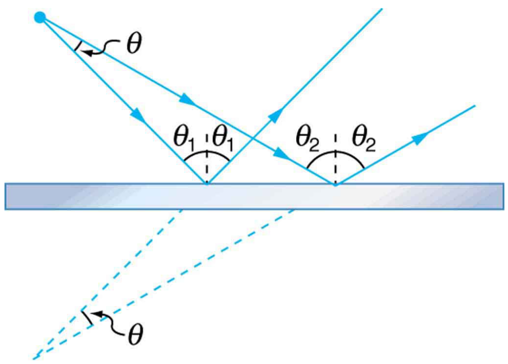
Problem:

Light shows staged with lasers use moving mirrors to swing beams and create colorful effects. Show that a light ray reflected from a mirror changes direction by 2θ when the mirror is rotated by an angle θ.

Exercise:

Problem:

A flat mirror is neither converging nor diverging. To prove this, consider two rays originating from the same point and diverging at an angle θ. Show that after striking a plane mirror, the angle between their directions remains θ.



A flat mirror neither converges nor diverges light rays. Two rays continue to diverge at the same angle after reflection.

Glossary

mirror

smooth surface that reflects light at specific angles, forming an image of the person or object in front of it

law of reflection
angle of reflection equals the angle of incidence

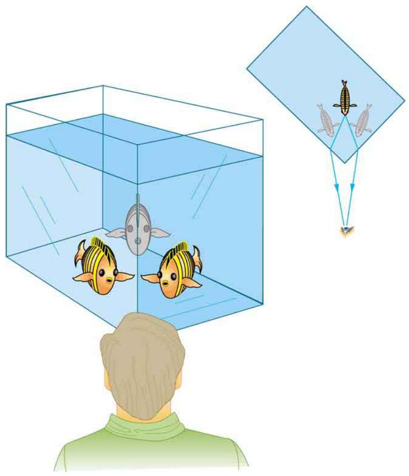
The Law of Refraction

- Determine the index of refraction, given the speed of light in a medium.

It is easy to notice some odd things when looking into a fish tank. For example, you may see the same fish appearing to be in two different places. (See [link].) This is because light coming from the fish to us changes direction when it leaves the tank, and in this case, it can travel two different paths to get to our eyes. The changing of a light ray’s direction (loosely called bending) when it passes through variations in matter is called **refraction**. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to voice transmission through optical fibers.

Note:
Refraction
The changing of a light ray’s direction (loosely called bending) when it passes through variations in matter is called refraction.

Note:
Speed of Light
The speed of light c not only affects refraction, it is one of the central concepts of Einstein’s theory of relativity. As the accuracy of the measurements of the speed of light were improved, c was found not to depend on the velocity of the source or the observer. However, the speed of light does vary in a precise manner with the material it traverses. These facts have far-reaching implications, as we will see in Special Relativity. It makes connections between space and time and alters our expectations that all observers measure the same time for the same event, for example. The speed of light is so important that its value in a vacuum is one of the most fundamental constants in nature as well as being one of the four fundamental SI units.



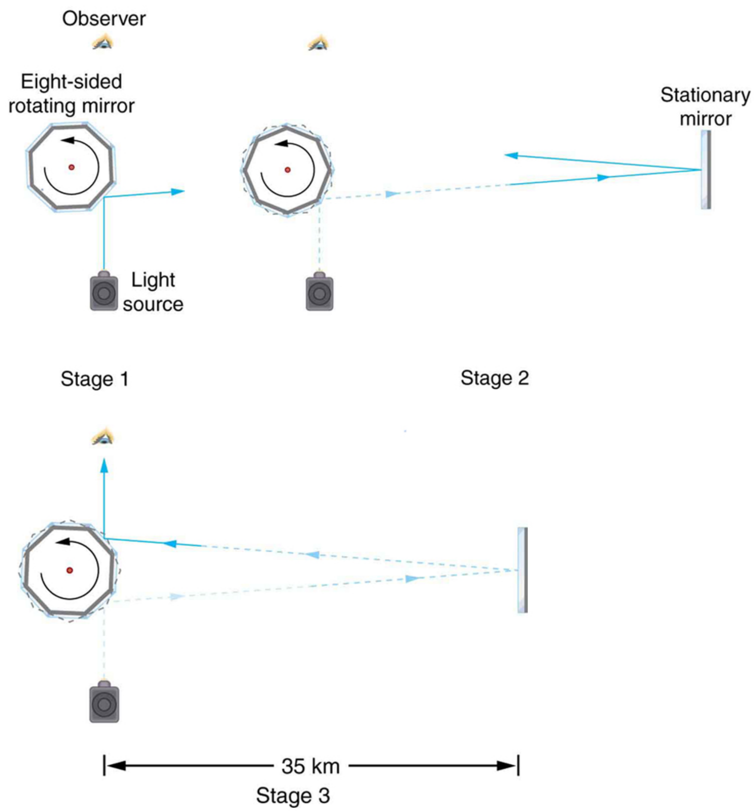
Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, and so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena.

Why does light change direction when passing from one material (medium) to another? It is because light changes speed when going from one material

to another. So before we study the law of refraction, it is useful to discuss the speed of light and how it varies in different media.

The Speed of Light

Early attempts to measure the speed of light, such as those made by Galileo, determined that light moved extremely fast, perhaps instantaneously. The first real evidence that light traveled at a finite speed came from the Danish astronomer Ole Roemer in the late 17th century. Roemer had noted that the average orbital period of one of Jupiter’s moons, as measured from Earth, varied depending on whether Earth was moving toward or away from Jupiter. He correctly concluded that the apparent change in period was due to the change in distance between Earth and Jupiter and the time it took light to travel this distance. From his 1676 data, a value of the speed of light was calculated to be $2.26 \times 10^8 \text{ m/s}$ (only 25% different from today’s accepted value). In more recent times, physicists have measured the speed of light in numerous ways and with increasing accuracy. One particularly direct method, used in 1887 by the American physicist Albert Michelson (1852–1931), is illustrated in [link]. Light reflected from a rotating set of mirrors was reflected from a stationary mirror 35 km away and returned to the rotating mirrors. The time for the light to travel can be determined by how fast the mirrors must rotate for the light to be returned to the observer’s eye.



A schematic of early apparatus used by Michelson and others to determine the speed of light. As the mirrors rotate, the reflected ray is only briefly directed at the stationary mirror. The returning ray will be reflected into the observer's eye only if the next mirror has rotated into the correct position just as the ray returns. By measuring the correct rotation rate, the time for the round trip can be measured and the speed of light calculated. Michelson’s calculated value of the speed of light was only 0.04% different from the value used today.

The speed of light is now known to great precision. In fact, the speed of light in a vacuum c is so important that it is accepted as one of the basic physical quantities and has the fixed value

Equation:

$$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s},$$

where the approximate value of $3.00 \times 10^8 \text{ m/s}$ is used whenever three-digit accuracy is sufficient. The speed of light through matter is less than it is in a vacuum, because light interacts with atoms in a material. The speed of light depends strongly on the type of material, since its interaction with different atoms, crystal lattices, and other substructures varies. We define the **index of refraction** n of a material to be

Equation:

$$n = \frac{c}{v},$$

where v is the observed speed of light in the material. Since the speed of light is always less than c in matter and equals c only in a vacuum, the index of refraction is always greater than or equal to one.

Note:
Value of the Speed of Light
Equation:

$$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$$

Note:
Index of Refraction
Equation:

$$n = \frac{c}{v}$$

That is, $n \geq 1$. [link] gives the indices of refraction for some representative substances. The values are listed for a particular wavelength of light, because they vary slightly with wavelength. (This can have important effects, such as colors produced by a prism.) Note that for gases, n is close to 1.0. This seems reasonable, since atoms in gases are widely separated and light travels at c in the vacuum between atoms. It is common to take $n = 1$ for gases unless great precision is needed. Although the speed of light v in a medium varies considerably from its value c in a vacuum, it is still a large speed.

Medium	***n***
Gases at $0^{\circ}\mathrm{C}$***, 1 atm***	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at $20^{\circ}\mathrm{C}$	
Benzene	1.501
Carbon disulfide	1.628

Medium	***n***
Carbon tetrachloride	1.461
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at 20°C	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice at 20°C	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

Index of Refraction in Various Media

Example:
Speed of Light in Matter
Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.
Strategy
The speed of light in a material, v, can be calculated from the index of refraction n of the material using the equation $n = c/v$.
Solution
The equation for index of refraction states that $n = c/v$. Rearranging this to determine v gives
Equation:

$$v = \frac{c}{n}.$$

The index of refraction for zircon is given as 1.923 in [link], and c is given in the equation for speed of light. Entering these values in the last expression gives
Equation:

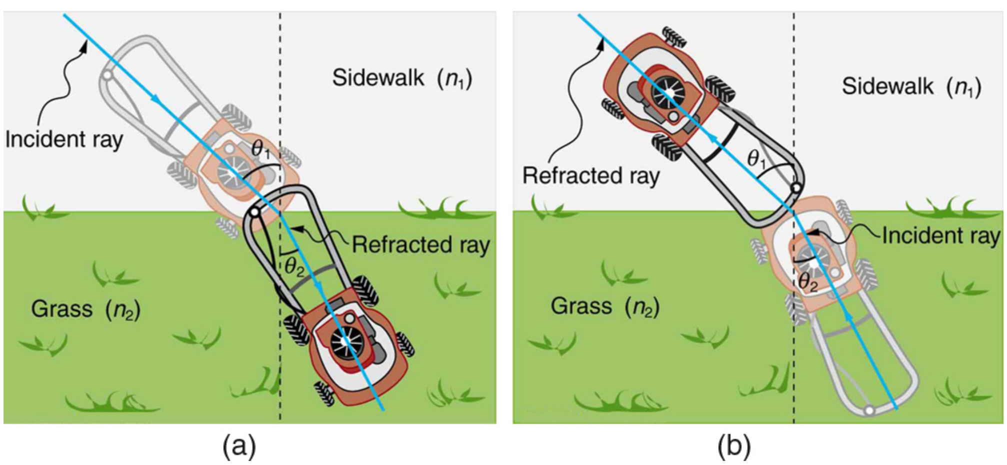
$$\begin{aligned} v &= \frac{3.00 \times 10^8 \text{ m/s}}{1.923} \\ &= 1.56 \times 10^8 \text{ m/s}. \end{aligned}$$

Discussion
This speed is slightly larger than half the speed of light in a vacuum and is still high compared with speeds we normally experience. The only substance listed in [link] that has a greater index of refraction than zircon is diamond. We shall see later that the large index of refraction for zircon makes it sparkle more than glass, but less than diamond.

Law of Refraction

[link] shows how a ray of light changes direction when it passes from one medium to another. As before, the angles are measured relative to a perpendicular to the surface at the point where the light ray crosses it.

(Some of the incident light will be reflected from the surface, but for now we will concentrate on the light that is transmitted.) The change in direction of the light ray depends on how the speed of light changes. The change in the speed of light is related to the indices of refraction of the media involved. In the situations shown in [link], medium 2 has a greater index of refraction than medium 1. This means that the speed of light is less in medium 2 than in medium 1. Note that as shown in [link](a), the direction of the ray moves closer to the perpendicular when it slows down. Conversely, as shown in [link](b), the direction of the ray moves away from the perpendicular when it speeds up. The path is exactly reversible. In both cases, you can imagine what happens by thinking about pushing a lawn mower from a footpath onto grass, and vice versa. Going from the footpath to grass, the front wheels are slowed and pulled to the side as shown. This is the same change in direction as for light when it goes from a fast medium to a slow one. When going from the grass to the footpath, the front wheels can move faster and the mower changes direction as shown. This, too, is the same change in direction as for light going from slow to fast.



The change in direction of a light ray depends on how the speed of light changes when it crosses from one medium to another. The speed of light is greater in medium 1 than in medium 2 in the situations shown here. (a) A ray of light moves closer to the perpendicular when it slows down. This is analogous to what happens when a lawn mower goes from a footpath to grass. (b) A ray of

light moves away from the perpendicular when it speeds up. This is analogous to what happens when a lawn mower goes from grass to footpath. The paths are exactly reversible.

The amount that a light ray changes its direction depends both on the incident angle and the amount that the speed changes. For a ray at a given incident angle, a large change in speed causes a large change in direction, and thus a large change in angle. The exact mathematical relationship is the **law of refraction**, or “Snell’s Law,” which is stated in equation form as
Equation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2.$$

Here n_1 and n_2 are the indices of refraction for medium 1 and 2, and θ_1 and θ_2 are the angles between the rays and the perpendicular in medium 1 and 2, as shown in [link]. The incoming ray is called the incident ray and the outgoing ray the refracted ray, and the associated angles the incident angle and the refracted angle. The law of refraction is also called Snell’s law after the Dutch mathematician Willebrord Snell (1591–1626), who discovered it in 1621. Snell’s experiments showed that the law of refraction was obeyed and that a characteristic index of refraction n could be assigned to a given medium. Snell was not aware that the speed of light varied in different media, but through experiments he was able to determine indices of refraction from the way light rays changed direction.

Note:
The Law of Refraction
Equation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

Note:
Take-Home Experiment: A Broken Pencil
A classic observation of refraction occurs when a pencil is placed in a glass half filled with water. Do this and observe the shape of the pencil when you look at the pencil sideways, that is, through air, glass, water. Explain your observations. Draw ray diagrams for the situation.

Example:
Determine the Index of Refraction from Refraction Data
Find the index of refraction for medium 2 in [link](a), assuming medium 1 is air and given the incident angle is 30.0° and the angle of refraction is 22.0°.
Strategy
The index of refraction for air is taken to be 1 in most cases (and up to four significant figures, it is 1.000). Thus $n_1 = 1.00$ here. From the given information, $\theta_1 = 30.0^{\circ}$ and $\theta_2 = 22.0^{\circ}$. With this information, the only unknown in Snell's law is n_2, so that it can be used to find this unknown.
Solution
Snell's law is
Equation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2.$$

Rearranging to isolate n_2 gives
Equation:

$$n_2 = n_1 \frac{\sin \theta_1}{\sin \theta_2}.$$

Entering known values,
Equation:

$$\begin{aligned} n_2 &= 1.00 \frac{\sin 30.0^{\circ}}{\sin 22.0^{\circ}} = \frac{0.500}{0.375} \\ &= 1.33. \end{aligned}$$

Discussion
This is the index of refraction for water, and Snell could have determined it by measuring the angles and performing this calculation. He would then have found 1.33 to be the appropriate index of refraction for water in all other situations, such as when a ray passes from water to glass. Today we can verify that the index of refraction is related to the speed of light in a medium by measuring that speed directly.

Example:
A Larger Change in Direction
Suppose that in a situation like that in [link], light goes from air to diamond and that the incident angle is 30.0°. Calculate the angle of refraction θ_2 in the diamond.
Strategy
Again the index of refraction for air is taken to be $n_1 = 1.00$, and we are given $\theta_1 = 30.0^{\circ}$. We can look up the index of refraction for diamond in [link], finding $n_2 = 2.419$. The only unknown in Snell's law is θ_2, which we wish to determine.
Solution
Solving Snell's law for sin θ_2 yields
Equation:

$$\sin\,\theta_2 = \frac{n_1}{n_2}\sin\,\theta_1.$$

Entering known values,
Equation:

$$\sin\,\theta_2 = \frac{1.00}{2.419}\sin\,30.0^{\circ} = \left(0.413\right)(0.500) = 0.207.$$

The angle is thus
Equation:

$$\theta_2 = \sin^{-1}0.207 = 11.9^{\circ}.$$

Discussion
For the same 30^{o} angle of incidence, the angle of refraction in diamond is significantly smaller than in water (11.9^{o} rather than 22^{o}—see the preceding example). This means there is a larger change in direction in diamond. The cause of a large change in direction is a large change in the index of refraction (or speed). In general, the larger the change in speed, the greater the effect on the direction of the ray.

Section Summary

- The changing of a light ray's direction when it passes through variations in matter is called refraction.
- The speed of light in vacuum $c = 2.99792458 \times 10^8\ \mathrm{m/s} \approx 3.00 \times 10^8\ \mathrm{m/s}.$
- Index of refraction $n = \frac{c}{v}$, where v is the speed of light in the material, c is the speed of light in vacuum, and n is the index of refraction.
- Snell's law, the law of refraction, is stated in equation form as $n_1 \sin \theta_1 = n_2 \sin \theta_2$.

Conceptual Questions

Exercise:

Problem:

Diffusion by reflection from a rough surface is described in this chapter. Light can also be diffused by refraction. Describe how this occurs in a specific situation, such as light interacting with crushed ice.

Exercise:

Problem:

Why is the index of refraction always greater than or equal to 1?

Exercise:

Problem:

Does the fact that the light flash from lightning reaches you before its sound prove that the speed of light is extremely large or simply that it is greater than the speed of sound? Discuss how you could use this effect to get an estimate of the speed of light.

Exercise:

Problem:

Will light change direction toward or away from the perpendicular when it goes from air to water? Water to glass? Glass to air?

Exercise:

Problem:

Explain why an object in water always appears to be at a depth shallower than it actually is? Why do people sometimes sustain neck and spinal injuries when diving into unfamiliar ponds or waters?

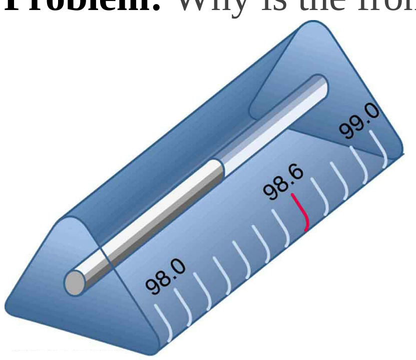
Exercise:

Problem:

Explain why a person's legs appear very short when wading in a pool. Justify your explanation with a ray diagram showing the path of rays from the feet to the eye of an observer who is out of the water.

Exercise:

Problem: Why is the front surface of a thermometer curved as shown?



The curved surface
of the thermometer
serves a purpose.

Exercise:

Problem:

Suppose light were incident from air onto a material that had a negative index of refraction, say –1.3; where does the refracted light ray go?

Problems & Exercises

Exercise:

Problem: What is the speed of light in water? In glycerine?

Solution:

$2.25 \times 10^8\ \mathrm{m/s}$ in water

$2.04 \times 10^8\ \mathrm{m/s}$ in glycerine

Exercise:

Problem: What is the speed of light in air? In crown glass?

Exercise:

Problem:

Calculate the index of refraction for a medium in which the speed of light is $2.012 \times 10^8\ \mathrm{m/s}$, and identify the most likely substance based on [link].

Solution:

1.490, polystyrene

Exercise:

Problem:

In what substance in [link] is the speed of light $2.290 \times 10^{8}\ \text{m/s}$?

Exercise:

Problem:

There was a major collision of an asteroid with the Moon in medieval times. It was described by monks at Canterbury Cathedral in England as a red glow on and around the Moon. How long after the asteroid hit the Moon, which is $3.84 \times 10^{5}\ \text{km}$ away, would the light first arrive on Earth?

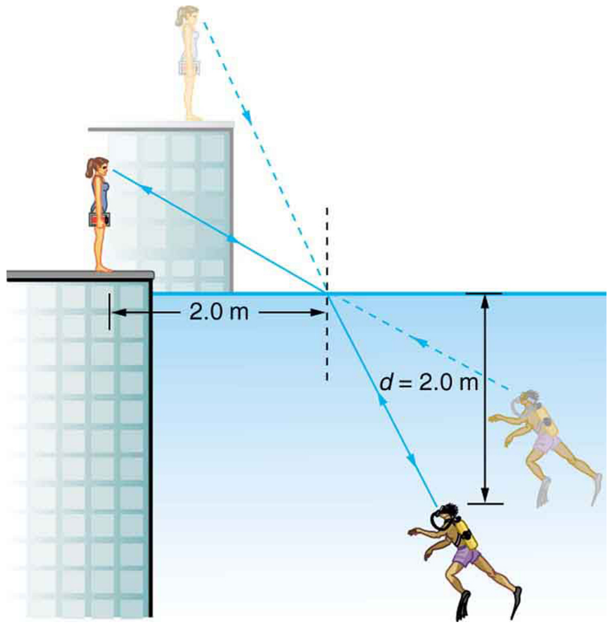
Solution:

$1.28\ \text{s}$

Exercise:

Problem:

A scuba diver training in a pool looks at his instructor as shown in [link]. What angle does the ray from the instructor's face make with the perpendicular to the water at the point where the ray enters? The angle between the ray in the water and the perpendicular to the water is 25.0°.



A scuba diver in a pool and his trainer look at each other.

Exercise:

Problem:

Components of some computers communicate with each other through optical fibers having an index of refraction $n = 1.55$. What time in nanoseconds is required for a signal to travel 0.200 m through such a fiber?

Solution:

1.03 ns

Exercise:

Problem:

(a) Given that the angle between the ray in the water and the perpendicular to the water is 25.0°, and using information in [link], find the height of the instructor's head above the water, noting that you will first have to calculate the angle of incidence. (b) Find the apparent depth of the diver's head below water as seen by the instructor.

Exercise:

Problem:

Suppose you have an unknown clear substance immersed in water, and you wish to identify it by finding its index of refraction. You arrange to have a beam of light enter it at an angle of 45.0°, and you observe the angle of refraction to be 40.3°. What is the index of refraction of the substance and its likely identity?

Solution:

$n = 1.46$, fused quartz

Exercise:

Problem:

On the Moon's surface, lunar astronauts placed a corner reflector, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. What percent correction is needed to account for the delay in time due to the slowing of light in Earth's atmosphere? Assume the distance to the Moon is precisely $3.84 \times 10^{8}\ \text{m}$, and Earth's atmosphere (which varies in density with altitude) is equivalent to a layer 30.0 km thick with a constant index of refraction $n = 1.000293$.

Exercise:

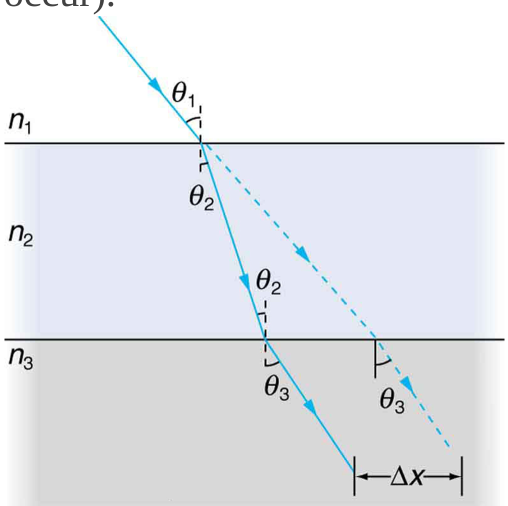
Problem:

Suppose [link] represents a ray of light going from air through crown glass into water, such as going into a fish tank. Calculate the amount the ray is displaced by the glass (Δx), given that the incident angle is 40.0º and the glass is 1.00 cm thick.

Exercise:

Problem:

[link] shows a ray of light passing from one medium into a second and then a third. Show that θ_3 is the same as it would be if the second medium were not present (provided total internal reflection does not occur).



A ray of light passes from one medium to a third by traveling through a second. The final direction is the same as if the second medium were not present, but the ray is displaced by Δx (shown exaggerated).

Exercise:

Problem: **Unreasonable Results**

Suppose light travels from water to another substance, with an angle of incidence of 10.0° and an angle of refraction of 14.9°. (a) What is the index of refraction of the other substance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) 0.898

(b) Can't have $n < 1.00$ since this would imply a speed greater than c.

(c) Refracted angle is too big relative to the angle of incidence.

Exercise:

Problem: **Construct Your Own Problem**

Consider sunlight entering the Earth's atmosphere at sunrise and sunset —that is, at a 90° incident angle. Taking the boundary between nearly empty space and the atmosphere to be sudden, calculate the angle of refraction for sunlight. This lengthens the time the Sun appears to be above the horizon, both at sunrise and sunset. Now construct a problem in which you determine the angle of refraction for different models of the atmosphere, such as various layers of varying density. Your instructor may wish to guide you on the level of complexity to consider and on how the index of refraction varies with air density.

Exercise:

Problem: **Unreasonable Results**

Light traveling from water to a gemstone strikes the surface at an angle of 80.0° and has an angle of refraction of 15.2°. (a) What is the speed

of light in the gemstone? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) $\frac{c}{5.00}$

(b) Speed of light too slow, since index is much greater than that of diamond.

(c) Angle of refraction is unreasonable relative to the angle of incidence.

Glossary

refraction
: changing of a light ray’s direction when it passes through variations in matter

index of refraction
: for a material, the ratio of the speed of light in vacuum to that in the material

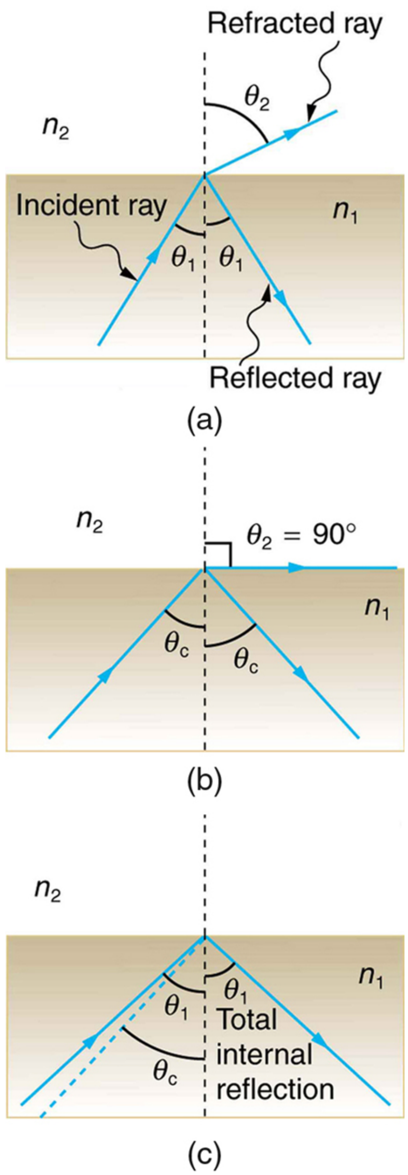
Total Internal Reflection

- Explain the phenomenon of total internal reflection.
- Describe the workings and uses of fiber optics.
- Analyze the reason for the sparkle of diamonds.

A good-quality mirror may reflect more than 90% of the light that falls on it, absorbing the rest. But it would be useful to have a mirror that reflects all of the light that falls on it. Interestingly, we can produce *total reflection* using an aspect of *refraction*.

Consider what happens when a ray of light strikes the surface between two materials, such as is shown in [link](a). Part of the light crosses the boundary and is refracted; the rest is reflected. If, as shown in the figure, the index of refraction for the second medium is less than for the first, the ray bends away from the perpendicular. (Since $n_1 > n_2$, the angle of refraction is greater than the angle of incidence—that is, $\theta_2 > \theta_1$.) Now imagine what happens as the incident angle is increased. This causes θ_2 to increase also. The largest the angle of refraction $\theta_2$can be is 90º, as shown in [link](b).The **critical angle**θ_c for a combination of materials is defined to be the incident angle θ_1 that produces an angle of refraction of 90º. That is, θ_c is the incident angle for which $\theta_2 = 90º$. If the incident angle θ_1 is greater than the critical angle, as shown in [link](c), then all of the light is reflected back into medium 1, a condition called **total internal reflection**.

Note:
Critical Angle
The incident angle θ_1 that produces an angle of refraction of 90º is called the critical angle, θ_c.



(a) A ray of light crosses a boundary where the speed of light increases and the index of refraction decreases. That is, $n_2 < n_1$. The ray bends away from the perpendicular. (b) The critical

angle θ_c is the one for which the angle of refraction is . (c) Total internal reflection occurs when the incident angle is greater than the critical angle.

Snell's law states the relationship between angles and indices of refraction. It is given by

Equation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2.$$

When the incident angle equals the critical angle ($\theta_1 = \theta_c$), the angle of refraction is 90º ($\theta_2 = 90º$). Noting that sin 90º=1, Snell's law in this case becomes

Equation:

$$n_1 \sin \theta_1 = n_2.$$

The critical angle θ_c for a given combination of materials is thus

Equation:

$$\theta_c = \sin^{-1}(n_2/n_1) \text{ for } n_1 > n_2.$$

Total internal reflection occurs for any incident angle greater than the critical angle θ_c, and it can only occur when the second medium has an index of refraction less than the first. Note the above equation is written for a light ray that travels in medium 1 and reflects from medium 2, as shown in the figure.

Example:
How Big is the Critical Angle Here?
What is the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air?
Strategy
The index of refraction for polystyrene is found to be 1.49 in [link], and the index of refraction of air can be taken to be 1.00, as before. Thus, the condition that the second medium (air) has an index of refraction less than the first (plastic) is satisfied, and the equation $\theta_c = \sin^{-1}(n_2/n_1)$ can be used to find the critical angle θ_c. Here, then, $n_2 = 1.00$ and $n_1 = 1.49$.
Solution
The critical angle is given by
Equation:

$$\theta_c = \sin^{-1}(n_2/n_1).$$

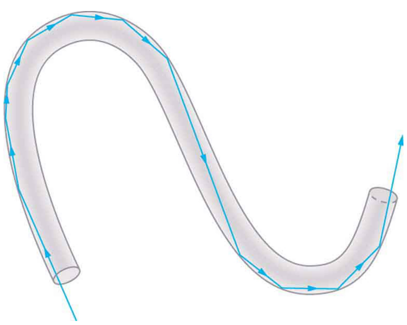
Substituting the identified values gives
Equation:

$$\theta_c = \sin^{-1}(1.00/1.49) = \sin^{-1}(0.671)$$
$$42.2^{\text{o}}.$$

Discussion
This means that any ray of light inside the plastic that strikes the surface at an angle greater than 42.2^{o} will be totally reflected. This will make the inside surface of the clear plastic a perfect mirror for such rays without any need for the silvering used on common mirrors. Different combinations of materials have different critical angles, but any combination with $n_1 > n_2$ can produce total internal reflection. The same calculation as made here shows that the critical angle for a ray going from water to air is 48.6^{o}, while that from diamond to air is 24.4^{o}, and that from flint glass to crown glass is 66.3^{o}. There is no total reflection for rays going in the other direction—for example, from air to water—since the condition that the second medium must have a smaller index of refraction is not satisfied. A number of interesting applications of total internal reflection follow.

Fiber Optics: Endoscopes to Telephones

Fiber optics is one application of total internal reflection that is in wide use. In communications, it is used to transmit telephone, internet, and cable TV signals. **Fiber optics** employs the transmission of light down fibers of plastic or glass. Because the fibers are thin, light entering one is likely to strike the inside surface at an angle greater than the critical angle and, thus, be totally reflected (See [link].) The index of refraction outside the fiber must be smaller than inside, a condition that is easily satisfied by coating the outside of the fiber with a material having an appropriate refractive index. In fact, most fibers have a varying refractive index to allow more light to be guided along the fiber through total internal refraction. Rays are reflected around corners as shown, making the fibers into tiny light pipes.

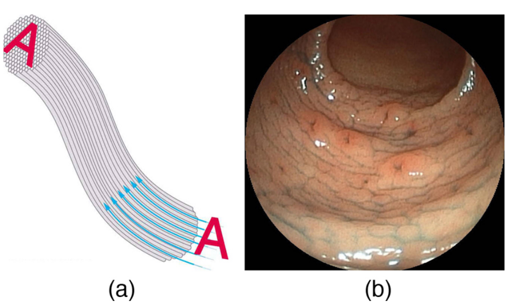


Light entering a thin fiber may strike the inside surface at large or grazing angles and is completely reflected if these angles exceed the critical angle. Such rays continue down the fiber, even following it around corners, since the angles of reflection

and incidence
remain large.

Bundles of fibers can be used to transmit an image without a lens, as illustrated in [link]. The output of a device called an **endoscope** is shown in [link](b). Endoscopes are used to explore the body through various orifices or minor incisions. Light is transmitted down one fiber bundle to illuminate internal parts, and the reflected light is transmitted back out through another to be observed. Surgery can be performed, such as arthroscopic surgery on the knee joint, employing cutting tools attached to and observed with the endoscope. Samples can also be obtained, such as by lassoing an intestinal polyp for external examination.

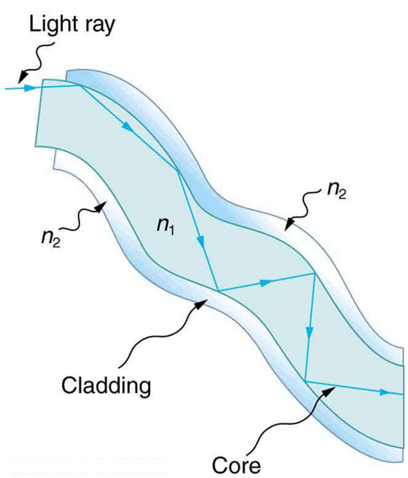
Fiber optics has revolutionized surgical techniques and observations within the body. There are a host of medical diagnostic and therapeutic uses. The flexibility of the fiber optic bundle allows it to navigate around difficult and small regions in the body, such as the intestines, the heart, blood vessels, and joints. Transmission of an intense laser beam to burn away obstructing plaques in major arteries as well as delivering light to activate chemotherapy drugs are becoming commonplace. Optical fibers have in fact enabled microsurgery and remote surgery where the incisions are small and the surgeon’s fingers do not need to touch the diseased tissue.



(a) An image is
transmitted by a bundle of
fibers that have fixed

neighbors. (b) An endoscope is used to probe the body, both transmitting light to the interior and returning an image such as the one shown. (credit: Med_Chaos, Wikimedia Commons)

Fibers in bundles are surrounded by a cladding material that has a lower index of refraction than the core. (See [link].) The cladding prevents light from being transmitted between fibers in a bundle. Without cladding, light could pass between fibers in contact, since their indices of refraction are identical. Since no light gets into the cladding (there is total internal reflection back into the core), none can be transmitted between clad fibers that are in contact with one another. The cladding prevents light from escaping out of the fiber; instead most of the light is propagated along the length of the fiber, minimizing the loss of signal and ensuring that a quality image is formed at the other end. The cladding and an additional protective layer make optical fibers flexible and durable.



Fibers in bundles are clad by a material that has a lower index of refraction than the core to ensure total internal reflection, even when fibers are in contact with one another. This shows a single fiber with its cladding.

Note:
Cladding
The cladding prevents light from being transmitted between fibers in a bundle.

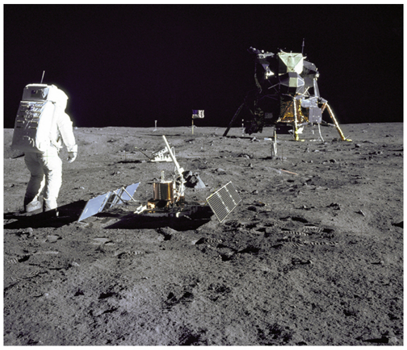
Special tiny lenses that can be attached to the ends of bundles of fibers are being designed and fabricated. Light emerging from a fiber bundle can be focused and a tiny spot can be imaged. In some cases the spot can be scanned, allowing quality imaging of a region inside the body. Special minute optical filters inserted at the end of the fiber bundle have the capacity to image tens of microns below the surface without cutting the surface—non-intrusive diagnostics. This is particularly useful for determining the extent of cancers in the stomach and bowel.

Most telephone conversations and Internet communications are now carried by laser signals along optical fibers. Extensive optical fiber cables have been placed on the ocean floor and underground to enable optical communications. Optical fiber communication systems offer several advantages over electrical (copper) based systems, particularly for long

distances. The fibers can be made so transparent that light can travel many kilometers before it becomes dim enough to require amplification—much superior to copper conductors. This property of optical fibers is called *low loss*. Lasers emit light with characteristics that allow far more conversations in one fiber than are possible with electric signals on a single conductor. This property of optical fibers is called *high bandwidth*. Optical signals in one fiber do not produce undesirable effects in other adjacent fibers. This property of optical fibers is called *reduced crosstalk*. We shall explore the unique characteristics of laser radiation in a later chapter.

Corner Reflectors and Diamonds

A light ray that strikes an object consisting of two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came. This is true whenever the reflecting surfaces are perpendicular, and it is independent of the angle of incidence. Such an object, shown in [link], is called a **corner reflector**, since the light bounces from its inside corner. Many inexpensive reflector buttons on bicycles, cars, and warning signs have corner reflectors designed to return light in the direction from which it originated. It was more expensive for astronauts to place one on the moon. Laser signals can be bounced from that corner reflector to measure the gradually increasing distance to the moon with great precision.

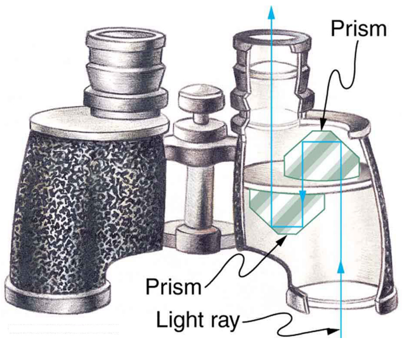


(a)

(b)

(a) Astronauts placed a corner reflector on the moon to measure its gradually increasing orbital distance. (credit: NASA) (b) The bright spots on these bicycle safety reflectors are reflections of the flash of the camera that took this picture on a dark night. (credit: Julo, Wikimedia Commons)

Corner reflectors are perfectly efficient when the conditions for total internal reflection are satisfied. With common materials, it is easy to obtain a critical angle that is less than 45°. One use of these perfect mirrors is in binoculars, as shown in [link]. Another use is in periscopes found in submarines.

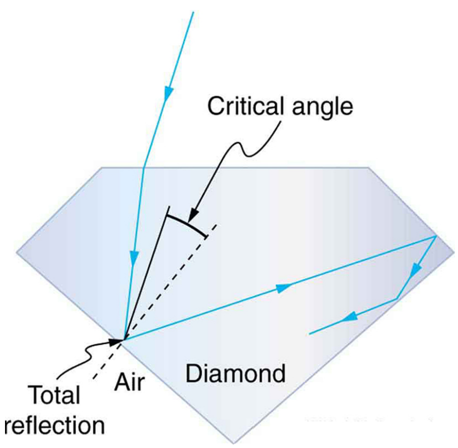


These binoculars employ corner reflectors with total internal reflection to get light to the observer's eyes.

The Sparkle of Diamonds

Total internal reflection, coupled with a large index of refraction, explains why diamonds sparkle more than other materials. The critical angle for a diamond-to-air surface is only 24.4°, and so when light enters a diamond, it has trouble getting back out. (See [link].) Although light freely enters the diamond, it can exit only if it makes an angle less than 24.4°. Facets on diamonds are specifically intended to make this unlikely, so that the light can exit only in certain places. Good diamonds are very clear, so that the light makes many internal reflections and is concentrated at the few places it can exit—hence the sparkle. (Zircon is a natural gemstone that has an exceptionally large index of refraction, but not as large as diamond, so it is

not as highly prized. Cubic zirconia is manufactured and has an even higher index of refraction (≈ 2.17), but still less than that of diamond.) The colors you see emerging from a sparkling diamond are not due to the diamond's color, which is usually nearly colorless. Those colors result from dispersion, the topic of Dispersion: The Rainbow and Prisms. Colored diamonds get their color from structural defects of the crystal lattice and the inclusion of minute quantities of graphite and other materials. The Argyle Mine in Western Australia produces around 90% of the world's pink, red, champagne, and cognac diamonds, while around 50% of the world's clear diamonds come from central and southern Africa.



Light cannot easily escape a diamond, because its critical angle with air is so small. Most reflections are total, and the facets are placed so that light can exit only in particular ways—thus concentrating the light and making the diamond sparkle.

Note:
PhET Explorations: Bending Light
Explore bending of light between two media with different indices of refraction. See how changing from air to water to glass changes the bending angle. Play with prisms of different shapes and make rainbows.

https://phet.colorado.edu/sims/html/bending-light/latest/bending-light_en.html

Section Summary

- The incident angle that produces an angle of refraction of 90° is called critical angle.
- Total internal reflection is a phenomenon that occurs at the boundary between two mediums, such that if the incident angle in the first medium is greater than the critical angle, then all the light is reflected back into that medium.
- Fiber optics involves the transmission of light down fibers of plastic or glass, applying the principle of total internal reflection.
- Endoscopes are used to explore the body through various orifices or minor incisions, based on the transmission of light through optical fibers.
- Cladding prevents light from being transmitted between fibers in a bundle.
- Diamonds sparkle due to total internal reflection coupled with a large index of refraction.

Conceptual Questions

Exercise:

Problem:

A ring with a colorless gemstone is dropped into water. The gemstone becomes invisible when submerged. Can it be a diamond? Explain.

Exercise:

Problem:

A high-quality diamond may be quite clear and colorless, transmitting all visible wavelengths with little absorption. Explain how it can sparkle with flashes of brilliant color when illuminated by white light.

Exercise:

Problem:

Is it possible that total internal reflection plays a role in rainbows? Explain in terms of indices of refraction and angles, perhaps referring to [link]. Some of us have seen the formation of a double rainbow. Is it physically possible to observe a triple rainbow?

Double rainbows are not a very common observance. (credit: InvictusOU812, Flickr)

Exercise:

Problem:

The most common type of mirage is an illusion that light from faraway objects is reflected by a pool of water that is not really there. Mirages are generally observed in deserts, when there is a hot layer of air near the ground. Given that the refractive index of air is lower for air at higher temperatures, explain how mirages can be formed.

Problems & Exercises

Exercise:

Problem:

Verify that the critical angle for light going from water to air is 48.6°, as discussed at the end of [link], regarding the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air.

Exercise:

Problem:

(a) At the end of [link], it was stated that the critical angle for light going from diamond to air is 24.4°. Verify this. (b) What is the critical angle for light going from zircon to air?

Exercise:

Problem:

An optical fiber uses flint glass clad with crown glass. What is the critical angle?

Solution:

66.3°

Exercise:

Problem:

At what minimum angle will you get total internal reflection of light traveling in water and reflected from ice?

Exercise:

Problem:

Suppose you are using total internal reflection to make an efficient corner reflector. If there is air outside and the incident angle is 45.0°, what must be the minimum index of refraction of the material from which the reflector is made?

Solution:

> 1.414

Exercise:

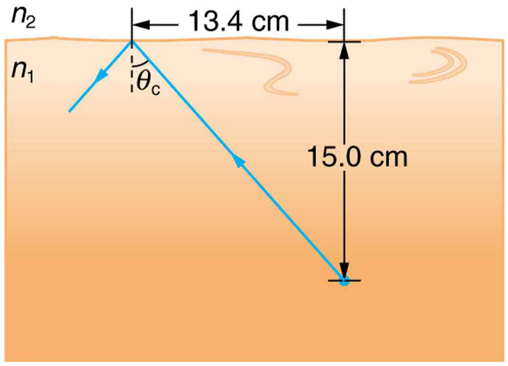
Problem:

You can determine the index of refraction of a substance by determining its critical angle. (a) What is the index of refraction of a substance that has a critical angle of 68.4° when submerged in water? What is the substance, based on [link]? (b) What would the critical angle be for this substance in air?

Exercise:

Problem:

A ray of light, emitted beneath the surface of an unknown liquid with air above it, undergoes total internal reflection as shown in [link]. What is the index of refraction for the liquid and its likely identification?



A light ray inside a liquid strikes the surface at the critical angle and undergoes total internal reflection.

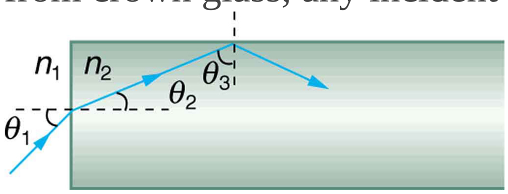
Solution:

1.50, benzene

Exercise:

Problem:

A light ray entering an optical fiber surrounded by air is first refracted and then reflected as shown in [link]. Show that if the fiber is made from crown glass, any incident ray will be totally internally reflected.



A light ray enters the end of a fiber, the surface of which is perpendicular to its sides. Examine the conditions under which it

may be totally internally reflected.

Glossary

critical angle
: incident angle that produces an angle of refraction of 90°

fiber optics
: transmission of light down fibers of plastic or glass, applying the principle of total internal reflection

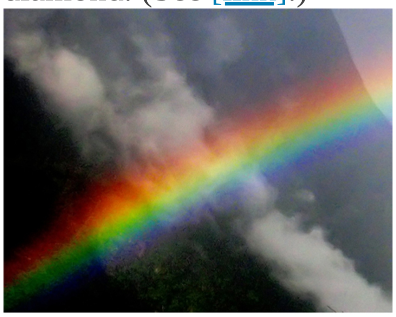
corner reflector
: an object consisting of two mutually perpendicular reflecting surfaces, so that the light that enters is reflected back exactly parallel to the direction from which it came

zircon
: natural gemstone with a large index of refraction

Dispersion: The Rainbow and Prisms

- Explain the phenomenon of dispersion and discuss its advantages and disadvantages.

Everyone enjoys the spectacle of a rainbow glimmering against a dark stormy sky. How does sunlight falling on clear drops of rain get broken into the rainbow of colors we see? The same process causes white light to be broken into colors by a clear glass prism or a diamond. (See [link].)



(a)

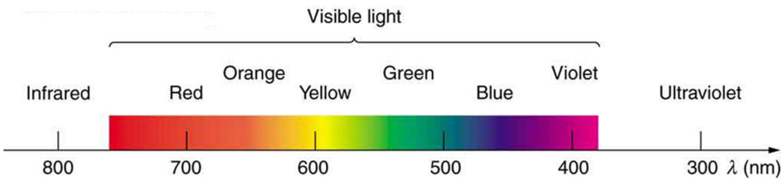
(b)

The colors of the rainbow (a) and those produced by a prism (b) are identical. (credit: Alfredo55, Wikimedia Commons; NASA)

We see about six colors in a rainbow—red, orange, yellow, green, blue, and violet; sometimes indigo is listed, too. Those colors are associated with different wavelengths of light, as shown in [link]. When our eye receives pure-wavelength light, we tend to see only one of the six colors, depending on wavelength. The thousands of other hues we can sense in other situations are our eye’s response to various mixtures of wavelengths. White light, in particular, is a fairly uniform mixture of all visible wavelengths. Sunlight, considered to be white, actually appears to be a bit yellow because of its mixture of wavelengths, but it does contain all visible wavelengths. The sequence of colors in rainbows is the same sequence as the colors plotted versus wavelength in [link]. What this implies is that white light is spread out according to

wavelength in a rainbow. **Dispersion** is defined as the spreading of white light into its full spectrum of wavelengths. More technically, dispersion occurs whenever there is a process that changes the direction of light in a manner that depends on wavelength. Dispersion, as a general phenomenon, can occur for any type of wave and always involves wavelength-dependent processes.

Note:
Dispersion
Dispersion is defined to be the spreading of white light into its full spectrum of wavelengths.



Even though rainbows are associated with seven colors, the rainbow is a continuous distribution of colors according to wavelengths.

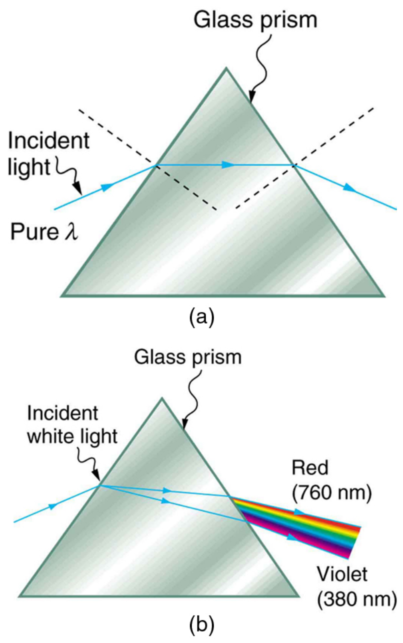
Refraction is responsible for dispersion in rainbows and many other situations. The angle of refraction depends on the index of refraction, as we saw in The Law of Refraction. We know that the index of refraction n depends on the medium. But for a given medium, n also depends on wavelength. (See [link]. Note that, for a given medium, n increases as wavelength decreases and is greatest for violet light. Thus violet light is bent more than red light, as shown for a prism in [link](b), and the light is dispersed into the same sequence of wavelengths as seen in [link] and [link].

Note:
Making Connections: Dispersion
Any type of wave can exhibit dispersion. Sound waves, all types of electromagnetic waves, and water waves can be dispersed according to wavelength. Dispersion occurs whenever the speed of propagation depends on wavelength, thus separating and spreading out various wavelengths. Dispersion may require special circumstances and can result in spectacular displays such as in the production of a rainbow. This is also

true for sound, since all frequencies ordinarily travel at the same speed. If you listen to sound through a long tube, such as a vacuum cleaner hose, you can easily hear it is dispersed by interaction with the tube. Dispersion, in fact, can reveal a great deal about what the wave has encountered that disperses its wavelengths. The dispersion of electromagnetic radiation from outer space, for example, has revealed much about what exists between the stars—the so-called empty space.

Medium	**Red (660 nm)**	**Orange (610 nm)**	**Yellow (580 nm)**	**Green (550 nm)**	**Blue (470 nm)**	**Violet (410 nm)**
Water	1.331	1.332	1.333	1.335	1.338	1.342
Diamond	2.410	2.415	2.417	2.426	2.444	2.458
Glass, crown	1.512	1.514	1.518	1.519	1.524	1.530
Glass, flint	1.662	1.665	1.667	1.674	1.684	1.698
Polystyrene	1.488	1.490	1.492	1.493	1.499	1.506
Quartz, fused	1.455	1.456	1.458	1.459	1.462	1.468

Index of Refraction n in Selected Media at Various Wavelengths

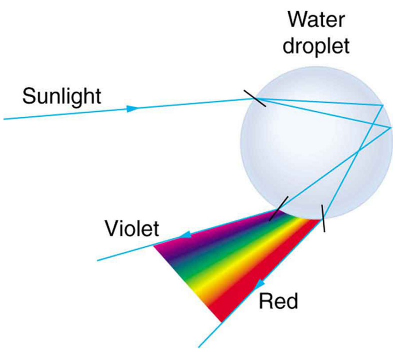


(a) A pure wavelength of light falls onto a prism and is refracted at both surfaces. (b) White light is dispersed by the prism (shown exaggerated). Since the index of refraction varies with wavelength, the angles of refraction vary with wavelength. A sequence of red to violet is produced, because the index of refraction increases steadily with decreasing wavelength.

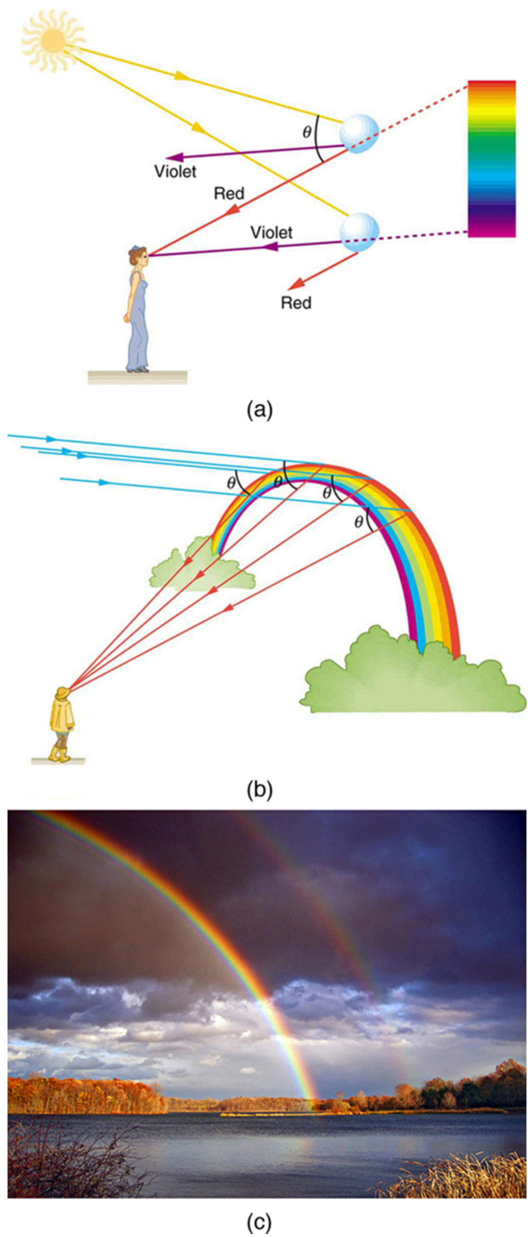
Rainbows are produced by a combination of refraction and reflection. You may have noticed that you see a rainbow only when you look away from the sun. Light enters a drop of water and is reflected from the back of the drop, as shown in [link]. The light is refracted both as it enters and as it leaves the drop. Since the index of refraction of water

varies with wavelength, the light is dispersed, and a rainbow is observed, as shown in [link] (a). (There is no dispersion caused by reflection at the back surface, since the law of reflection does not depend on wavelength.) The actual rainbow of colors seen by an observer depends on the myriad of rays being refracted and reflected toward the observer's eyes from numerous drops of water. The effect is most spectacular when the background is dark, as in stormy weather, but can also be observed in waterfalls and lawn sprinklers. The arc of a rainbow comes from the need to be looking at a specific angle relative to the direction of the sun, as illustrated in [link] (b). (If there are two reflections of light within the water drop, another "secondary" rainbow is produced. This rare event produces an arc that lies above the primary rainbow arc—see [link] (c).)

Note:
Rainbows
Rainbows are produced by a combination of refraction and reflection.



Part of the light falling on this water drop enters and is reflected from the back of the drop. This light is refracted and dispersed both as it enters and as it leaves the drop.



(a) Different colors emerge in different directions, and so you must look at different locations to see the various colors of a rainbow. (b) The arc of a rainbow results from the fact that a line between the observer and any point on the arc must make the correct angle with the parallel rays of sunlight to receive the refracted rays. (c)

Double rainbow. (credit: Nicholas, Wikimedia Commons)

Dispersion may produce beautiful rainbows, but it can cause problems in optical systems. White light used to transmit messages in a fiber is dispersed, spreading out in time and eventually overlapping with other messages. Since a laser produces a nearly pure wavelength, its light experiences little dispersion, an advantage over white light for transmission of information. In contrast, dispersion of electromagnetic waves coming to us from outer space can be used to determine the amount of matter they pass through. As with many phenomena, dispersion can be useful or a nuisance, depending on the situation and our human goals.

Note:
PhET Explorations: Geometric Optics
How does a lens form an image? See how light rays are refracted by a lens. Watch how the image changes when you adjust the focal length of the lens, move the object, move the lens, or move the screen.

https://phet.colorado.edu/sims/geometric-optics/geometric-optics_en.html

Section Summary

- The spreading of white light into its full spectrum of wavelengths is called dispersion.
- Rainbows are produced by a combination of refraction and reflection and involve the dispersion of sunlight into a continuous distribution of colors.
- Dispersion produces beautiful rainbows but also causes problems in certain optical systems.

Problems & Exercises

Exercise:

Problem:

(a) What is the ratio of the speed of red light to violet light in diamond, based on [link]? (b) What is this ratio in polystyrene? (c) Which is more dispersive?

Exercise:

Problem:

A beam of white light goes from air into water at an incident angle of 75.0°. At what angles are the red (660 nm) and violet (410 nm) parts of the light refracted?

Solution:

46.5°, red; 46.0°, violet

Exercise:

Problem:

By how much do the critical angles for red (660 nm) and violet (410 nm) light differ in a diamond surrounded by air?

Exercise:

Problem:

(a) A narrow beam of light containing yellow (580 nm) and green (550 nm) wavelengths goes from polystyrene to air, striking the surface at a 30.0° incident angle. What is the angle between the colors when they emerge? (b) How far would they have to travel to be separated by 1.00 mm?

Solution:

(a) 0.043°

(b) 1.33 m

Exercise:

Problem:

A parallel beam of light containing orange (610 nm) and violet (410 nm) wavelengths goes from fused quartz to water, striking the surface between them at a 60.0° incident angle. What is the angle between the two colors in water?

Exercise:

Problem:

A ray of 610 nm light goes from air into fused quartz at an incident angle of 55.0°. At what incident angle must 470 nm light enter flint glass to have the same angle of refraction?

Solution:

71.3°

Exercise:

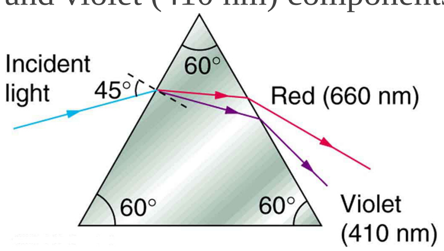
Problem:

A narrow beam of light containing red (660 nm) and blue (470 nm) wavelengths travels from air through a 1.00 cm thick flat piece of crown glass and back to air again. The beam strikes at a 30.0° incident angle. (a) At what angles do the two colors emerge? (b) By what distance are the red and blue separated when they emerge?

Exercise:

Problem:

A narrow beam of white light enters a prism made of crown glass at a 45.0° incident angle, as shown in [link]. At what angles, θ_{R} and θ_{V}, do the red (660 nm) and violet (410 nm) components of the light emerge from the prism?



This prism will disperse the white light into a rainbow of colors. The incident angle is 45.0°, and the angles at which the red and violet light emerge are θ_{R} and θ_{V}.

Solution:

53.5°, red; 55.2°, violet

Glossary

dispersion

spreading of white light into its full spectrum of wavelengths

rainbow

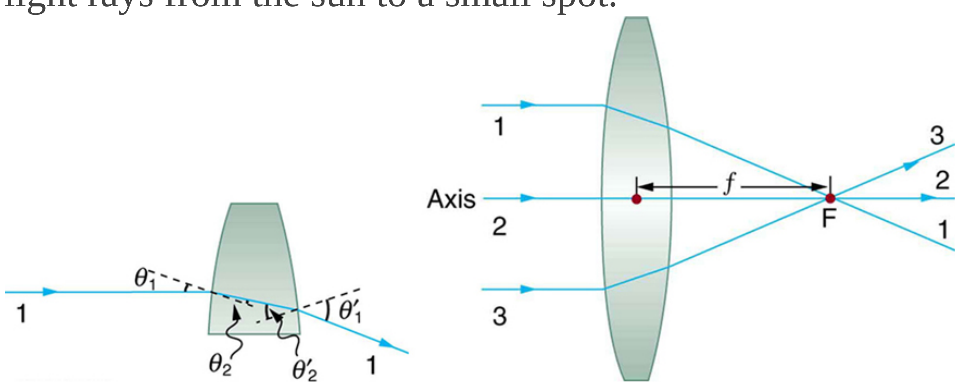
dispersion of sunlight into a continuous distribution of colors according to wavelength, produced by the refraction and reflection of sunlight by water droplets in the sky

Image Formation by Lenses

- List the rules for ray tracking for thin lenses.
- Illustrate the formation of images using the technique of ray tracking.
- Determine power of a lens given the focal length.

Lenses are found in a huge array of optical instruments, ranging from a simple magnifying glass to the eye to a camera's zoom lens. In this section, we will use the law of refraction to explore the properties of lenses and how they form images.

The word *lens* derives from the Latin word for a lentil bean, the shape of which is similar to the convex lens in [link]. The convex lens shown has been shaped so that all light rays that enter it parallel to its axis cross one another at a single point on the opposite side of the lens. (The axis is defined to be a line normal to the lens at its center, as shown in [link].) Such a lens is called a **converging (or convex) lens** for the converging effect it has on light rays. An expanded view of the path of one ray through the lens is shown, to illustrate how the ray changes direction both as it enters and as it leaves the lens. Since the index of refraction of the lens is greater than that of air, the ray moves towards the perpendicular as it enters and away from the perpendicular as it leaves. (This is in accordance with the law of refraction.) Due to the lens's shape, light is thus bent toward the axis at both surfaces. The point at which the rays cross is defined to be the **focal point** F of the lens. The distance from the center of the lens to its focal point is defined to be the **focal length** f of the lens. [link] shows how a converging lens, such as that in a magnifying glass, can converge the nearly parallel light rays from the sun to a small spot.



Rays of light entering a converging lens parallel to its axis converge at its focal point F. (Ray 2 lies on the axis of the lens.) The distance from the center of the lens to the focal point is the lens's focal length f. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Note:
Converging or Convex Lens
The lens in which light rays that enter it parallel to its axis cross one another at a single point on the opposite side with a converging effect is called converging lens.

Note:
Focal Point F
The point at which the light rays cross is called the focal point F of the lens.

Note:
Focal Length f
The distance from the center of the lens to its focal point is called focal length f.

Sunlight focused by a converging magnifying glass can burn paper. Light rays from the sun are nearly parallel and cross at the focal point of the lens. The more powerful the lens, the closer to the lens the rays will cross.

The greater effect a lens has on light rays, the more powerful it is said to be. For example, a powerful converging lens will focus parallel light rays closer to itself and will have a smaller focal length than a weak lens. The light will also focus into a smaller and more intense spot for a more powerful lens. The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is

Equation:

$$P = \frac{1}{f}.$$

Note:
Power P
The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is
Equation:

$$P = \frac{1}{f}.$$

where f is the focal length of the lens, which must be given in meters (and not cm or mm). The power of a lens P has the unit diopters (D), provided that the focal length is given in meters. That is, $1\text{ D} = 1/\text{m}$, or 1 m^{-1}. (Note that this power (optical power, actually) is not the same as power in watts defined in Work, Energy, and Energy Resources. It is a concept related to the effect of optical devices on light.) Optometrists prescribe common spectacles and contact lenses in units of diopters.

Example:
What is the Power of a Common Magnifying Glass?
Suppose you take a magnifying glass out on a sunny day and you find that it concentrates sunlight to a small spot 8.00 cm away from the lens. What are the focal length and power of the lens?
Strategy
The situation here is the same as those shown in [link] and [link]. The Sun is so far away that the Sun's rays are nearly parallel when they reach Earth. The magnifying glass is a convex (or converging) lens, focusing the nearly parallel rays of sunlight. Thus the focal length of the lens is the distance from the lens to the spot, and its power is the inverse of this distance (in m).
Solution
The focal length of the lens is the distance from the center of the lens to the spot, given to be 8.00 cm. Thus,
Equation:

$$f = 8.00\text{ cm}.$$

To find the power of the lens, we must first convert the focal length to meters; then, we substitute this value into the equation for power. This gives

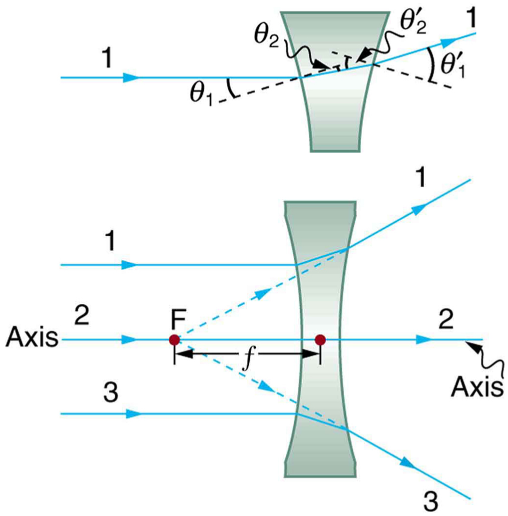
Equation:

$$P = \frac{1}{f} = \frac{1}{0.0800\ \text{m}} = 12.5\ \text{D}.$$

Discussion

This is a relatively powerful lens. The power of a lens in diopters should not be confused with the familiar concept of power in watts. It is an unfortunate fact that the word "power" is used for two completely different concepts. If you examine a prescription for eyeglasses, you will note lens powers given in diopters. If you examine the label on a motor, you will note energy consumption rate given as a power in watts.

[link] shows a concave lens and the effect it has on rays of light that enter it parallel to its axis (the path taken by ray 2 in the figure is the axis of the lens). The concave lens is a **diverging lens**, because it causes the light rays to bend away (diverge) from its axis. In this case, the lens has been shaped so that all light rays entering it parallel to its axis appear to originate from the same point, F, defined to be the focal point of a diverging lens. The distance from the center of the lens to the focal point is again called the focal length f of the lens. Note that the focal length and power of a diverging lens are defined to be negative. For example, if the distance to F in [link] is 5.00 cm, then the focal length is $f = –5.00\ \text{cm}$ and the power of the lens is $P = –20\ \text{D}$. An expanded view of the path of one ray through the lens is shown in the figure to illustrate how the shape of the lens, together with the law of refraction, causes the ray to follow its particular path and be diverged.

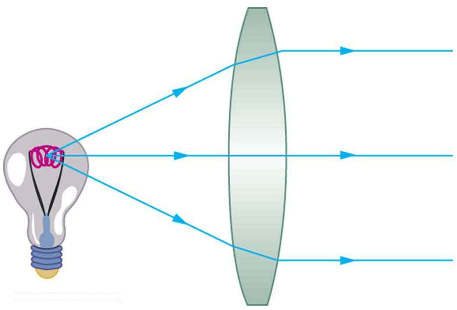


Rays of light entering a diverging lens parallel to its axis are diverged, and all appear to originate at its focal point F. The dashed lines are not rays—they indicate the directions from which the rays appear to come. The focal length f of a diverging lens is negative. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Note:
Diverging Lens

A lens that causes the light rays to bend away from its axis is called a diverging lens.

As noted in the initial discussion of the law of refraction in The Law of Refraction, the paths of light rays are exactly reversible. This means that the direction of the arrows could be reversed for all of the rays in [link] and [link]. For example, if a point light source is placed at the focal point of a convex lens, as shown in [link], parallel light rays emerge from the other side.



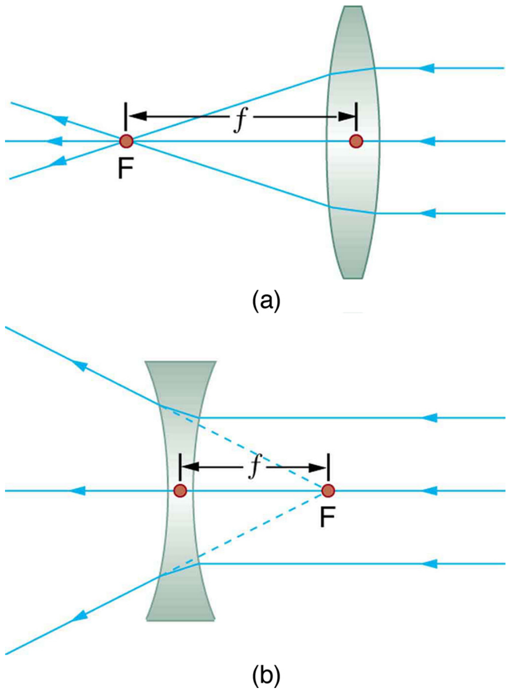
A small light source, like a light bulb filament, placed at the focal point of a convex lens, results in parallel rays of light emerging from the other side. The paths are exactly the reverse of those shown in [link]. This technique is used in lighthouses and sometimes in traffic lights to produce a directional beam of light from a source that emits light in all directions.

Ray Tracing and Thin Lenses

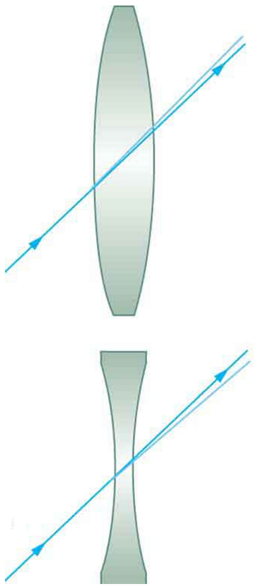
Ray tracing is the technique of determining or following (tracing) the paths that light rays take. For rays passing through matter, the law of refraction is used to trace the paths. Here we use ray tracing to help us understand the action of lenses in situations ranging from forming images on film to magnifying small print to correcting nearsightedness. While ray tracing for complicated lenses, such as those found in sophisticated cameras, may require computer techniques, there is a set of simple rules for tracing rays through thin lenses. A **thin lens** is defined to be one whose thickness allows rays to refract, as illustrated in [link], but does not allow properties such as dispersion and aberrations. An ideal thin lens has two refracting surfaces but the lens is thin enough to assume that light rays bend only once. A thin symmetrical lens has two focal points, one on either side and both at the same distance from the lens. (See [link].) Another important characteristic of a thin lens is that light rays through its center are deflected by a negligible amount, as seen in [link].

Note:
Thin Lens
A thin lens is defined to be one whose thickness allows rays to refract but does not allow properties such as dispersion and aberrations.

Note:
Take-Home Experiment: A Visit to the Optician
Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they act like thin lenses.



Thin lenses have the same focal length on either side. (a) Parallel light rays entering a converging lens from the right cross at its focal point on the left. (b) Parallel light rays entering a diverging lens from the right seem to come from the focal point on the right.



The light ray through the center of a thin lens is deflected by a negligible amount and is assumed to emerge parallel to its original path (shown as a shaded line).

Using paper, pencil, and a straight edge, ray tracing can accurately describe the operation of a lens. The rules for ray tracing for thin lenses are based on the illustrations already discussed:

1. A ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side. (See rays 1 and 3 in [link].)
2. A ray entering a diverging lens parallel to its axis seems to come from the focal point F. (See rays 1 and 3 in [link].)
3. A ray passing through the center of either a converging or a diverging lens does not change direction. (See [link], and see ray 2 in [link] and [link].)
4. A ray entering a converging lens through its focal point exits parallel to its axis. (The reverse of rays 1 and 3 in [link].)
5. A ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis. (The reverse of rays 1 and 3 in [link].)

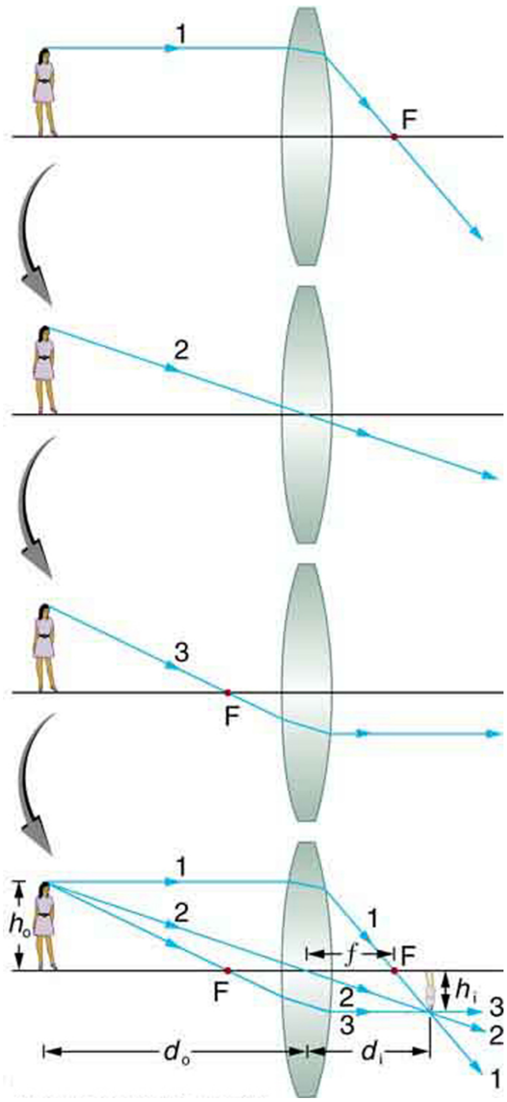
Note:
Rules for Ray Tracing

1. A ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side.
2. A ray entering a diverging lens parallel to its axis seems to come from the focal point F.
3. A ray passing through the center of either a converging or a diverging lens does not change direction.
4. A ray entering a converging lens through its focal point exits parallel to its axis.
5. A ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis.

Image Formation by Thin Lenses

In some circumstances, a lens forms an obvious image, such as when a movie projector casts an image onto a screen. In other cases, the image is less obvious. Where, for example, is the image formed by eyeglasses? We will use ray tracing for thin lenses to illustrate how they form images, and we will develop equations to describe the image formation quantitatively.

Consider an object some distance away from a converging lens, as shown in [link]. To find the location and size of the image formed, we trace the paths of selected light rays originating from one point on the object, in this case the top of the person’s head. The figure shows three rays from the top of the object that can be traced using the ray tracing rules given above. (Rays leave this point going in many directions, but we concentrate on only a few with paths that are easy to trace.) The first ray is one that enters the lens parallel to its axis and passes through the focal point on the other side (rule 1). The second ray passes through the center of the lens without changing direction (rule 3). The third ray passes through the nearer focal point on its way into the lens and leaves the lens parallel to its axis (rule 4). The three rays cross at the same point on the other side of the lens. The image of the top of the person’s head is located at this point. All rays that come from the same point on the top of the person’s head are refracted in such a way as to cross at the point shown. Rays from another point on the object, such as her belt buckle, will also cross at another common point, forming a complete image, as shown. Although three rays are traced in [link], only two are necessary to locate the image. It is best to trace rays for which there are simple ray tracing rules. Before applying ray tracing to other situations, let us consider the example shown in [link] in more detail.

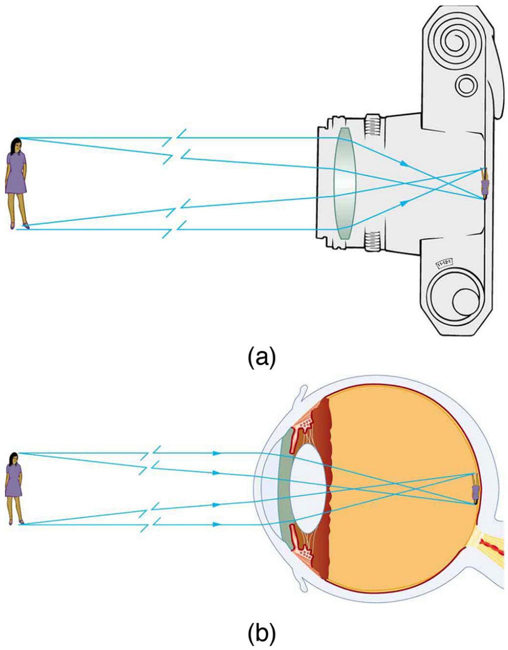


Ray tracing is used to locate the image formed by a lens. Rays originating from the same point on the object are traced—the three chosen rays each follow one of the rules for ray tracing, so that their paths are easy to determine. The image is located at the point where the rays

cross. In this case, a real image—one that can be projected on a screen—is formed.

The image formed in [link] is a **real image**, meaning that it can be projected. That is, light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye, for example. [link] shows how such an image would be projected onto film by a camera lens. This figure also shows how a real image is projected onto the retina by the lens of an eye. Note that the image is there whether it is projected onto a screen or not.

Note:
Real Image
The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.



Real images can be projected. (a) A real image of the person is projected onto film. (b) The converging nature of the multiple surfaces that make up the eye result in the projection of a real image on the retina.

Several important distances appear in [link]. We define d_o to be the object distance, the distance of an object from the center of a lens. **Image distance** d_i is defined to be the distance of the image from the center of a lens. The height of the object and height of the image are given the symbols h_o and h_i, respectively. Images that appear upright relative to the object have heights that are positive and those that are inverted have negative heights. Using the rules of ray tracing and making a scale drawing with paper and pencil, like that in [link], we can accurately describe the location and size of an image. But the real benefit of ray tracing is in visualizing how images are formed in a variety of situations. To obtain numerical information, we use a pair of

equations that can be derived from a geometric analysis of ray tracing for thin lenses. The **thin lens equations** are

Equation:

$$\frac{1}{d_{\mathrm{o}}}+\frac{1}{d_{\mathrm{i}}}=\frac{1}{f}$$

and

Equation:

$$\frac{h_{\mathrm{i}}}{h_{\mathrm{o}}}=-\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}}=m.$$

We define the ratio of image height to object height ($h_{\mathrm{i}}/h_{\mathrm{o}}$) to be the **magnification** m. (The minus sign in the equation above will be discussed shortly.) The thin lens equations are broadly applicable to all situations involving thin lenses (and “thin” mirrors, as we will see later). We will explore many features of image formation in the following worked examples.

Note:
Image Distance
The distance of the image from the center of the lens is called image distance.

Note:
Thin Lens Equations and Magnification
Equation:

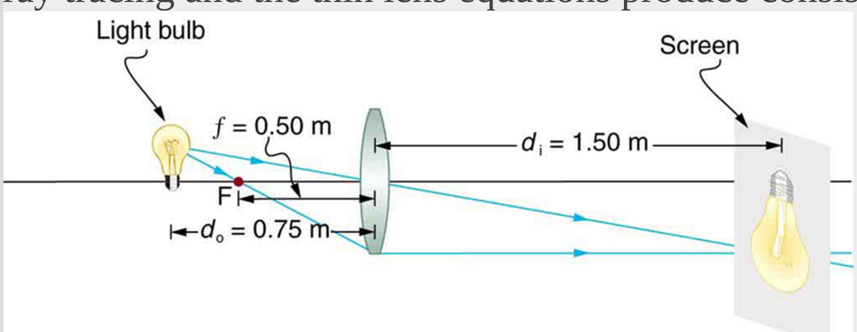
$$\frac{1}{d_{\mathrm{o}}}+\frac{1}{d_{\mathrm{i}}}=\frac{1}{f}$$

Equation:

$$\frac{h_{\rm i}}{h_{\rm o}} = -\frac{d_{\rm i}}{d_{\rm o}} = m$$

Example:
Finding the Image of a Light Bulb Filament by Ray Tracing and by the Thin Lens Equations

A clear glass light bulb is placed 0.750 m from a convex lens having a 0.500 m focal length, as shown in [link]. Use ray tracing to get an approximate location for the image. Then use the thin lens equations to calculate (a) the location of the image and (b) its magnification. Verify that ray tracing and the thin lens equations produce consistent results.



A light bulb placed 0.750 m from a lens having a 0.500 m focal length produces a real image on a poster board as discussed in the example above. Ray tracing predicts the image location and size.

Strategy and Concept

Since the object is placed farther away from a converging lens than the focal length of the lens, this situation is analogous to those illustrated in [link] and [link]. Ray tracing to scale should produce similar results for $d_{\rm i}$. Numerical solutions for $d_{\rm i}$ and m can be obtained using the thin lens equations, noting that $d_{\rm o} = 0.750\ {\rm m}$ and $f = 0.500\ {\rm m}$.

Solutions (Ray tracing)

The ray tracing to scale in [link] shows two rays from a point on the bulb's filament crossing about 1.50 m on the far side of the lens. Thus the image distance d_i is about 1.50 m. Similarly, the image height based on ray tracing is greater than the object height by about a factor of 2, and the image is inverted. Thus m is about –2. The minus sign indicates that the image is inverted.

The thin lens equations can be used to find d_i from the given information:

Equation:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}.$$

Rearranging to isolate d_i gives

Equation:

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}.$$

Entering known quantities gives a value for $1/d_i$:

Equation:

$$\frac{1}{d_i} = \frac{1}{0.500\ \text{m}} - \frac{1}{0.750\ \text{m}} = \frac{0.667}{\text{m}}.$$

This must be inverted to find d_i:

Equation:

$$d_i = \frac{\text{m}}{0.667} = 1.50\ \text{m}.$$

Note that another way to find d_i is to rearrange the equation:

Equation:

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}.$$

This yields the equation for the image distance as:

Equation:

$$d_{\mathrm{i}} = \frac{f d_{\mathrm{o}}}{d_{\mathrm{o}} - f}.$$

Note that there is no inverting here.
The thin lens equations can be used to find the magnification m, since both d_{i} and d_{o} are known. Entering their values gives
Equation:

$$m = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = -\frac{1.50\ \mathrm{m}}{0.750\ \mathrm{m}} = -2.00.$$

Discussion
Note that the minus sign causes the magnification to be negative when the image is inverted. Ray tracing and the use of the thin lens equations produce consistent results. The thin lens equations give the most precise results, being limited only by the accuracy of the given information. Ray tracing is limited by the accuracy with which you can draw, but it is highly useful both conceptually and visually.

Real images, such as the one considered in the previous example, are formed by converging lenses whenever an object is farther from the lens than its focal length. This is true for movie projectors, cameras, and the eye. We shall refer to these as *case 1* images. A case 1 image is formed when $d_{\mathrm{o}} > f$ and f is positive, as in [link](a). (A summary of the three cases or types of image formation appears at the end of this section.)

A different type of image is formed when an object, such as a person's face, is held close to a convex lens. The image is upright and larger than the object, as seen in [link](b), and so the lens is called a magnifier. If you slowly pull the magnifier away from the face, you will see that the magnification steadily increases until the image begins to blur. Pulling the magnifier even farther away produces an inverted image as seen in [link](a). The distance at which the image blurs, and beyond which it inverts, is the focal length of the lens. To use a convex lens as a magnifier, the object

must be closer to the converging lens than its focal length. This is called a *case 2* image. A case 2 image is formed when $d_o < f$ and f is positive.

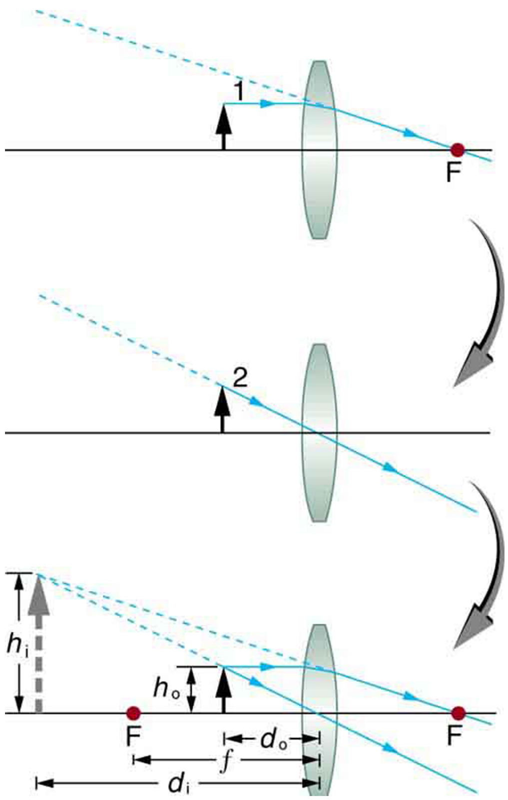
(a)

(b)

(a) When a converging lens is held farther away from the face than the lens's focal length, an inverted image is formed. This is a case 1 image. Note that the image is in focus but the face is not, because the image is much closer to the camera taking this photograph than the face. (credit: DaMongMan, Flickr) (b) A magnified image

of a face is produced by placing it closer to the converging lens than its focal length. This is a case 2 image. (credit: Casey Fleser, Flickr)

[link] uses ray tracing to show how an image is formed when an object is held closer to a converging lens than its focal length. Rays coming from a common point on the object continue to diverge after passing through the lens, but all appear to originate from a point at the location of the image. The image is on the same side of the lens as the object and is farther away from the lens than the object. This image, like all case 2 images, cannot be projected and, hence, is called a **virtual image**. Light rays only appear to originate at a virtual image; they do not actually pass through that location in space. A screen placed at the location of a virtual image will receive only diffuse light from the object, not focused rays from the lens. Additionally, a screen placed on the opposite side of the lens will receive rays that are still diverging, and so no image will be projected on it. We can see the magnified image with our eyes, because the lens of the eye converges the rays into a real image projected on our retina. Finally, we note that a virtual image is upright and larger than the object, meaning that the magnification is positive and greater than 1.



Ray tracing predicts the image location and size for an object held closer to a converging lens than its focal length. Ray 1 enters parallel to the axis and exits through the focal point on the opposite side, while ray 2 passes through the center of the lens without changing path. The two rays continue to diverge on the other side of the lens, but both appear to come from a common point, locating the upright, magnified,

virtual image. This is a
case 2 image.

Note:
Virtual Image
An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Example:
Image Produced by a Magnifying Glass
Suppose the book page in [link] (a) is held 7.50 cm from a convex lens of focal length 10.0 cm, such as a typical magnifying glass might have. What magnification is produced?
Strategy and Concept
We are given that $d_o = 7.50\text{ cm}$ and $f = 10.0\text{ cm}$, so we have a situation where the object is placed closer to the lens than its focal length. We therefore expect to get a case 2 virtual image with a positive magnification that is greater than 1. Ray tracing produces an image like that shown in [link], but we will use the thin lens equations to get numerical solutions in this example.
Solution
To find the magnification m, we try to use magnification equation, $m = -d_i/d_o$. We do not have a value for d_i, so that we must first find the location of the image using lens equation. (The procedure is the same as followed in the preceding example, where d_o and f were known.) Rearranging the magnification equation to isolate d_i gives
Equation:

$$\frac{1}{d_i} = \frac{1}{f} - \frac{1}{d_o}.$$

Entering known values, we obtain a value for $1/d_i$:
Equation:

$$\frac{1}{d_i} = \frac{1}{10.0\text{ cm}} - \frac{1}{7.50\text{ cm}} = \frac{-0.0333}{\text{cm}}.$$

This must be inverted to find d_i:

Equation:

$$d_i = -\frac{\text{cm}}{0.0333} = -30.0\text{ cm}.$$

Now the thin lens equation can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

Equation:

$$m = -\frac{d_i}{d_o} = -\frac{-30.0\text{ cm}}{7.50\text{ cm}} = 4.00.$$

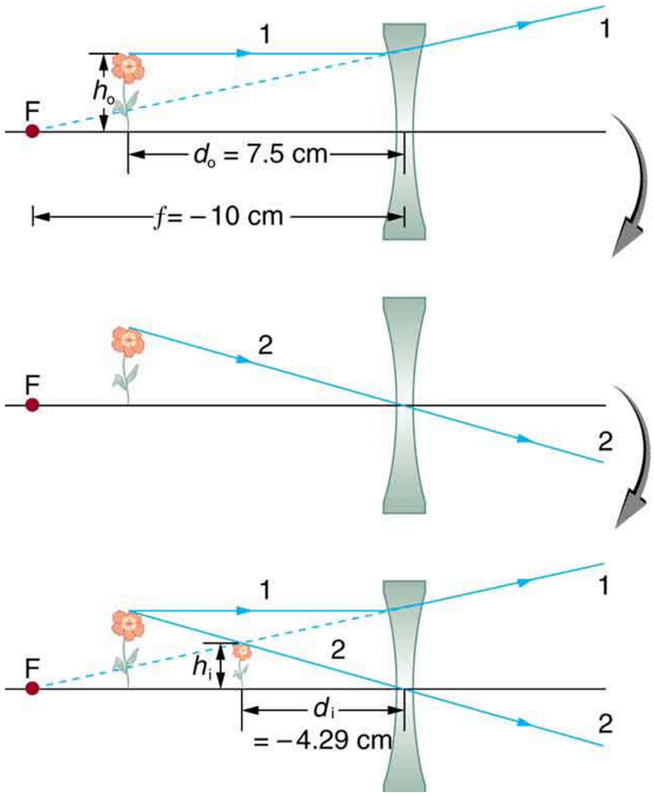
Discussion

A number of results in this example are true of all case 2 images, as well as being consistent with [link]. Magnification is indeed positive (as predicted), meaning the image is upright. The magnification is also greater than 1, meaning that the image is larger than the object—in this case, by a factor of 4. Note that the image distance is negative. This means the image is on the same side of the lens as the object. Thus the image cannot be projected and is virtual. (Negative values of d_i occur for virtual images.) The image is farther from the lens than the object, since the image distance is greater in magnitude than the object distance. The location of the image is not obvious when you look through a magnifier. In fact, since the image is bigger than the object, you may think the image is closer than the object. But the image is farther away, a fact that is useful in correcting farsightedness, as we shall see in a later section.

A third type of image is formed by a diverging or concave lens. Try looking through eyeglasses meant to correct nearsightedness. (See [link].) You will see an image that is upright but smaller than the object. This means that the magnification is positive but less than 1. The ray diagram in [link] shows that the image is on the same side of the lens as the object and, hence,

cannot be projected—it is a virtual image. Note that the image is closer to the lens than the object. This is a *case 3* image, formed for any object by a negative focal length or diverging lens.

A car viewed through a concave or diverging lens looks upright. This is a case 3 image. (credit: Daniel Oines, Flickr)



Ray tracing predicts the image location and size for a concave or diverging lens. Ray 1 enters parallel to the axis and is bent so that it appears to originate from the focal point. Ray 2 passes through the center of the lens without changing path. The two rays appear to come from a common point, locating the upright image. This is a case 3 image, which is closer to the lens than the object and smaller in height.

Example:

Image Produced by a Concave Lens

Suppose an object such as a book page is held 7.50 cm from a concave lens of focal length –10.0 cm. Such a lens could be used in eyeglasses to correct pronounced nearsightedness. What magnification is produced?

Strategy and Concept

This example is identical to the preceding one, except that the focal length is negative for a concave or diverging lens. The method of solution is thus the same, but the results are different in important ways.

Solution

To find the magnification m, we must first find the image distance d_{i} using thin lens equation

Equation:

$$\frac{1}{d_{\mathrm{i}}} = \frac{1}{f} - \frac{1}{d_{\mathrm{o}}},$$

or its alternative rearrangement

Equation:

$$d_i = \frac{f d_{\mathrm{o}}}{d_{\mathrm{o}} - f}.$$

We are given that $f = –10.0$ cm and $d_{\mathrm{o}} = 7.50$ cm. Entering these yields a value for $1/d_{\mathrm{i}}$:

Equation:

$$\frac{1}{d_{\mathrm{i}}} = \frac{1}{-10.0\ \mathrm{cm}} - \frac{1}{7.50\ \mathrm{cm}} = \frac{-0.2333}{\mathrm{cm}}.$$

This must be inverted to find d_{i}:

Equation:

$$d_{\mathrm{i}} = -\frac{\mathrm{cm}}{0.2333} = -4.29\ \mathrm{cm}.$$

Or

Equation:

$$d_i = \frac{(7.5)(-10)}{(7.5 - (-10))} = -75/17.5 = -4.29 \text{ cm}.$$

Now the magnification equation can be used to find the magnification m, since both d_i and d_o are known. Entering their values gives

Equation:

$$m = -\frac{d_i}{d_o} = -\frac{-4.29 \text{ cm}}{7.50 \text{ cm}} = 0.571.$$

Discussion

A number of results in this example are true of all case 3 images, as well as being consistent with [link]. Magnification is positive (as predicted), meaning the image is upright. The magnification is also less than 1, meaning the image is smaller than the object—in this case, a little over half its size. The image distance is negative, meaning the image is on the same side of the lens as the object. (The image is virtual.) The image is closer to the lens than the object, since the image distance is smaller in magnitude than the object distance. The location of the image is not obvious when you look through a concave lens. In fact, since the image is smaller than the object, you may think it is farther away. But the image is closer than the object, a fact that is useful in correcting nearsightedness, as we shall see in a later section.

[link] summarizes the three types of images formed by single thin lenses. These are referred to as case 1, 2, and 3 images. Convex (converging) lenses can form either real or virtual images (cases 1 and 2, respectively), whereas concave (diverging) lenses can form only virtual images (always case 3). Real images are always inverted, but they can be either larger or smaller than the object. For example, a slide projector forms an image larger than the slide, whereas a camera makes an image smaller than the object being photographed. Virtual images are always upright and cannot be projected. Virtual images are larger than the object only in case 2, where a convex lens is used. The virtual image produced by a concave lens is

always smaller than the object—a case 3 image. We can see and photograph virtual images only by using an additional lens to form a real image.

Type	**Formed when**	**Image type**	d_i	m
Case 1	f positive, $d_o > f$	real	positive	negative
Case 2	f positive, $d_o < f$	virtual	negative	positive $m > 1$
Case 3	f negative	virtual	negative	positive $m < 1$

Three Types of Images Formed By Thin Lenses

In Image Formation by Mirrors, we shall see that mirrors can form exactly the same types of images as lenses.

Note:

Take-Home Experiment: Concentrating Sunlight

Find several lenses and determine whether they are converging or diverging. In general those that are thicker near the edges are diverging and those that are thicker near the center are converging. On a bright sunny day take the converging lenses outside and try focusing the sunlight onto a piece of paper. Determine the focal lengths of the lenses. Be careful because the paper may start to burn, depending on the type of lens you have selected.

Problem-Solving Strategies for Lenses

Step 1. Examine the situation to determine that image formation by a lens is involved.

Step 2. Determine whether ray tracing, the thin lens equations, or both are to be employed. A sketch is very useful even if ray tracing is not specifically required by the problem. Write symbols and values on the sketch.

Step 3. Identify exactly what needs to be determined in the problem (identify the unknowns).

Step 4. Make alist of what is given or can be inferred from the problem as stated (identify the knowns). It is helpful to determine whether the situation involves a case 1, 2, or 3 image. While these are just names for types of images, they have certain characteristics (given in [link]) that can be of great use in solving problems.

Step 5. If ray tracing is required, use the ray tracing rules listed near the beginning of this section.

Step 6. Most quantitative problems require the use of the thin lens equations. These are solved in the usual manner by substituting knowns and solving for unknowns. Several worked examples serve as guides.

Step 7. Check to see if the answer is reasonable: Does it make sense? If you have identified the type of image (case 1, 2, or 3), you should assess whether your answer is consistent with the type of image, magnification, and so on.

Note:
Misconception Alert
We do not realize that light rays are coming from every part of the object, passing through every part of the lens, and all can be used to form the final image.
We generally feel the entire lens, or mirror, is needed to form an image. Actually, half a lens will form the same, though a fainter, image.

Section Summary

- Light rays entering a converging lens parallel to its axis cross one another at a single point on the opposite side.
- For a converging lens, the focal point is the point at which converging light rays cross; for a diverging lens, the focal point is the point from which diverging light rays appear to originate.
- The distance from the center of the lens to its focal point is called the focal length f.
- Power P of a lens is defined to be the inverse of its focal length, $P = \frac{1}{f}$.
- A lens that causes the light rays to bend away from its axis is called a diverging lens.
- Ray tracing is the technique of graphically determining the paths that light rays take.
- The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.
- Thin lens equations are $\frac{1}{d_{\mathrm{o}}} + \frac{1}{d_{\mathrm{i}}} = \frac{1}{f}$ and $\frac{h_{\mathrm{i}}}{h_{\mathrm{o}}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = m$ (magnification).

- The distance of the image from the center of the lens is called image distance.
- An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Conceptual Questions

Exercise:

Problem:

It can be argued that a flat piece of glass, such as in a window, is like a lens with an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?

Exercise:

Problem:

You can often see a reflection when looking at a sheet of glass, particularly if it is darker on the other side. Explain why you can often see a double image in such circumstances.

Exercise:

Problem:

When you focus a camera, you adjust the distance of the lens from the film. If the camera lens acts like a thin lens, why can it not be a fixed distance from the film for both near and distant objects?

Exercise:

Problem:

A thin lens has two focal points, one on either side, at equal distances from its center, and should behave the same for light entering from either side. Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they are thin lenses.

Exercise:

Problem:

Will the focal length of a lens change when it is submerged in water? Explain.

Problems & Exercises

Exercise:

Problem:

What is the power in diopters of a camera lens that has a 50.0 mm focal length?

Exercise:

Problem:

Your camera's zoom lens has an adjustable focal length ranging from 80.0 to 200 mm. What is its range of powers?

Solution:

$5.00 \text{ to } 12.5 \text{ D}$

Exercise:

Problem:

What is the focal length of 1.75 D reading glasses found on the rack in a pharmacy?

Exercise:

Problem:

You note that your prescription for new eyeglasses is –4.50 D. What will their focal length be?

Solution:

-0.222 m

Exercise:

Problem:

How far from the lens must the film in a camera be, if the lens has a 35.0 mm focal length and is being used to photograph a flower 75.0 cm away? Explicitly show how you follow the steps in the Problem-Solving Strategy for lenses.

Exercise:

Problem:

A certain slide projector has a 100 mm focal length lens. (a) How far away is the screen, if a slide is placed 103 mm from the lens and produces a sharp image? (b) If the slide is 24.0 by 36.0 mm, what are the dimensions of the image? Explicitly show how you follow the steps in the Problem-Solving Strategy for lenses.

Solution:

(a) 3.43 m

(b) 0.800 by 1.20 m

Exercise:

Problem:

A doctor examines a mole with a 15.0 cm focal length magnifying glass held 13.5 cm from the mole (a) Where is the image? (b) What is its magnification? (c) How big is the image of a 5.00 mm diameter mole?

Solution:

(a) -1.35 m (on the object side of the lens).

(b) $+10.0$

(c) 5.00 cm

Exercise:

Problem:

How far from a piece of paper must you hold your father's 2.25 D reading glasses to try to burn a hole in the paper with sunlight?

Solution:

44.4 cm

Exercise:

Problem:

A camera with a 50.0 mm focal length lens is being used to photograph a person standing 3.00 m away. (a) How far from the lens must the film be? (b) If the film is 36.0 mm high, what fraction of a 1.75 m tall person will fit on it? (c) Discuss how reasonable this seems, based on your experience in taking or posing for photographs.

Exercise:

Problem:

A camera lens used for taking close-up photographs has a focal length of 22.0 mm. The farthest it can be placed from the film is 33.0 mm. (a) What is the closest object that can be photographed? (b) What is the magnification of this closest object?

Solution:

(a) 6.60 cm

(b) –0.333

Exercise:

Problem:

Suppose your 50.0 mm focal length camera lens is 51.0 mm away from the film in the camera. (a) How far away is an object that is in focus? (b) What is the height of the object if its image is 2.00 cm high?

Exercise:

Problem:

(a) What is the focal length of a magnifying glass that produces a magnification of 3.00 when held 5.00 cm from an object, such as a rare coin? (b) Calculate the power of the magnifier in diopters. (c) Discuss how this power compares to those for store-bought reading glasses (typically 1.0 to 4.0 D). Is the magnifier's power greater, and should it be?

Solution:

(a) $+7.50 \text{ cm}$

(b) 13.3 D

(c) Much greater

Exercise:

Problem:

What magnification will be produced by a lens of power –4.00 D (such as might be used to correct myopia) if an object is held 25.0 cm away?

Exercise:

Problem:

In [link], the magnification of a book held 7.50 cm from a 10.0 cm focal length lens was found to be 3.00. (a) Find the magnification for the book when it is held 8.50 cm from the magnifier. (b) Do the same for when it is held 9.50 cm from the magnifier. (c) Comment on the trend in m as the object distance increases as in these two calculations.

Solution:

(a) +6.67

(b) +20.0

(c) The magnification increases without limit (to infinity) as the object distance increases to the limit of the focal distance.

Exercise:

Problem:

Suppose a 200 mm focal length telephoto lens is being used to photograph mountains 10.0 km away. (a) Where is the image? (b) What is the height of the image of a 1000 m high cliff on one of the mountains?

Exercise:

Problem:

A camera with a 100 mm focal length lens is used to photograph the sun and moon. What is the height of the image of the sun on the film, given the sun is 1.40×10^6 km in diameter and is 1.50×10^8 km away?

Solution:

-0.933 mm

Exercise:

Problem:

Combine thin lens equations to show that the magnification for a thin lens is determined by its focal length and the object distance and is given by $m = f/(f - d_\text{o})$.

Glossary

converging lens
: a convex lens in which light rays that enter it parallel to its axis converge at a single point on the opposite side

diverging lens
: a concave lens in which light rays that enter it parallel to its axis bend away (diverge) from its axis

focal point
: for a converging lens or mirror, the point at which converging light rays cross; for a diverging lens or mirror, the point from which diverging light rays appear to originate

focal length
: distance from the center of a lens or curved mirror to its focal point

magnification
: ratio of image height to object height

power
: inverse of focal length

real image
: image that can be projected

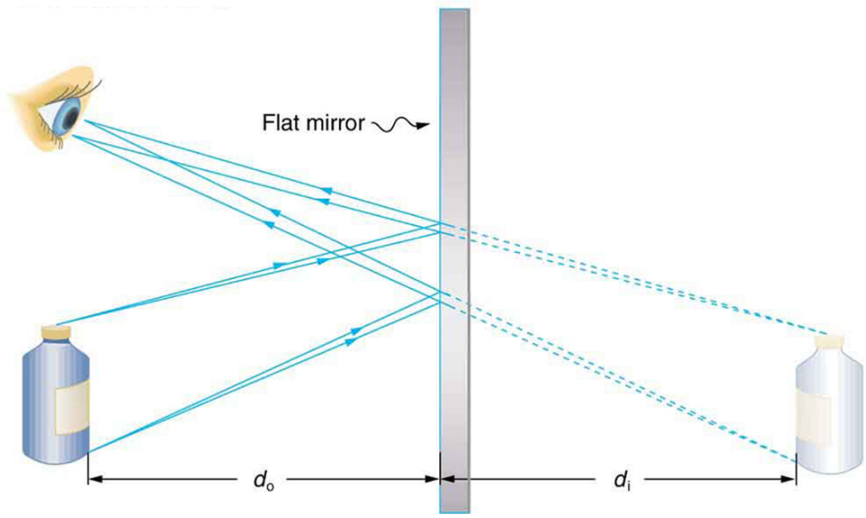
virtual image
: image that cannot be projected

Image Formation by Mirrors

- Illustrate image formation in a flat mirror.
- Explain with ray diagrams the formation of an image using spherical mirrors.
- Determine focal length and magnification given radius of curvature, distance of object and image.

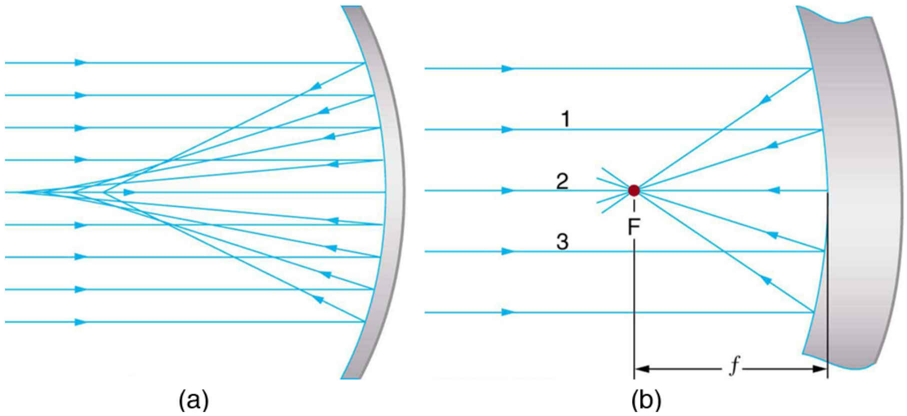
We only have to look as far as the nearest bathroom to find an example of an image formed by a mirror. Images in flat mirrors are the same size as the object and are located behind the mirror. Like lenses, mirrors can form a variety of images. For example, dental mirrors may produce a magnified image, just as makeup mirrors do. Security mirrors in shops, on the other hand, form images that are smaller than the object. We will use the law of reflection to understand how mirrors form images, and we will find that mirror images are analogous to those formed by lenses.

[link] helps illustrate how a flat mirror forms an image. Two rays are shown emerging from the same point, striking the mirror, and being reflected into the observer’s eye. The rays can diverge slightly, and both still get into the eye. If the rays are extrapolated backward, they seem to originate from a common point behind the mirror, locating the image. (The paths of the reflected rays into the eye are the same as if they had come directly from that point behind the mirror.) Using the law of reflection—the angle of reflection equals the angle of incidence—we can see that the image and object are the same distance from the mirror. This is a virtual image, since it cannot be projected—the rays only appear to originate from a common point behind the mirror. Obviously, if you walk behind the mirror, you cannot see the image, since the rays do not go there. But in front of the mirror, the rays behave exactly as if they had come from behind the mirror, so that is where the image is situated.



Two sets of rays from common points on an object are reflected by a flat mirror into the eye of an observer. The reflected rays seem to originate from behind the mirror, locating the virtual image.

Now let us consider the focal length of a mirror—for example, the concave spherical mirrors in [link]. Rays of light that strike the surface follow the law of reflection. For a mirror that is large compared with its radius of curvature, as in [link](a), we see that the reflected rays do not cross at the same point, and the mirror does not have a well-defined focal point. If the mirror had the shape of a parabola, the rays would all cross at a single point, and the mirror would have a well-defined focal point. But parabolic mirrors are much more expensive to make than spherical mirrors. The solution is to use a mirror that is small compared with its radius of curvature, as shown in [link](b). (This is the mirror equivalent of the thin lens approximation.) To a very good approximation, this mirror has a well-defined focal point at F that is the focal distance f from the center of the mirror. The focal length f of a concave mirror is positive, since it is a converging mirror.



(a) Parallel rays reflected from a large spherical mirror do not all cross at a common point. (b) If a spherical mirror is small compared with its radius of curvature, parallel rays are focused to a common point. The distance of the focal point from the center of the mirror is its focal length f. Since this mirror is converging, it has a positive focal length.

Just as for lenses, the shorter the focal length, the more powerful the mirror; thus, $P = 1/f$ for a mirror, too. A more strongly curved mirror has a shorter focal length and a greater power. Using the law of reflection and some simple trigonometry, it can be shown that the focal length is half the radius of curvature, or

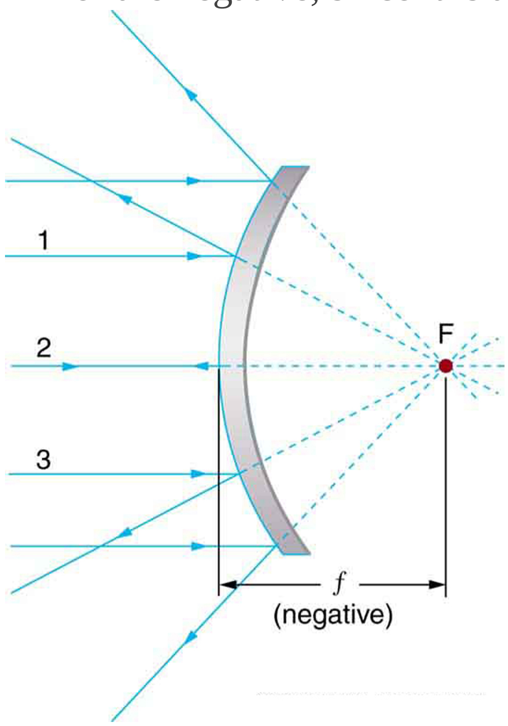
Equation:

$$f = \frac{R}{2},$$

where R is the radius of curvature of a spherical mirror. The smaller the radius of curvature, the smaller the focal length and, thus, the more powerful the mirror.

The convex mirror shown in [link] also has a focal point. Parallel rays of light reflected from the mirror seem to originate from the point F at the

focal distance f behind the mirror. The focal length and power of a convex mirror are negative, since it is a diverging mirror.



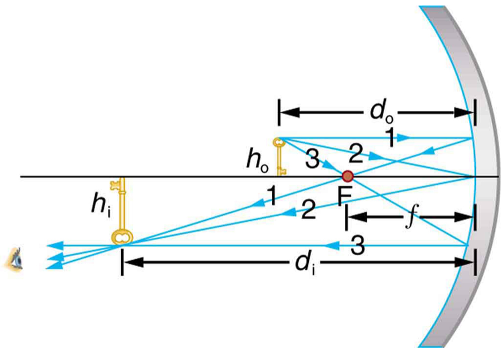
Parallel rays of light reflected from a convex spherical mirror (small in size compared with its radius of curvature) seem to originate from a well-defined focal point at the focal distance f behind the mirror. Convex mirrors diverge light rays and, thus, have a negative focal length.

Ray tracing is as useful for mirrors as for lenses. The rules for ray tracing for mirrors are based on the illustrations just discussed:

1. A ray approaching a concave converging mirror parallel to its axis is reflected through the focal point F of the mirror on the same side. (See rays 1 and 3 in [link](b).)
2. A ray approaching a convex diverging mirror parallel to its axis is reflected so that it seems to come from the focal point F behind the mirror. (See rays 1 and 3 in [link].)
3. Any ray striking the center of a mirror is followed by applying the law of reflection; it makes the same angle with the axis when leaving as when approaching. (See ray 2 in [link].)
4. A ray approaching a concave converging mirror through its focal point is reflected parallel to its axis. (The reverse of rays 1 and 3 in [link].)
5. A ray approaching a convex diverging mirror by heading toward its focal point on the opposite side is reflected parallel to the axis. (The reverse of rays 1 and 3 in [link].)

We will use ray tracing to illustrate how images are formed by mirrors, and we can use ray tracing quantitatively to obtain numerical information. But since we assume each mirror is small compared with its radius of curvature, we can use the thin lens equations for mirrors just as we did for lenses.

Consider the situation shown in [link], concave spherical mirror reflection, in which an object is placed farther from a concave (converging) mirror than its focal length. That is, f is positive and $d_o > f$, so that we may expect an image similar to the case 1 real image formed by a converging lens. Ray tracing in [link] shows that the rays from a common point on the object all cross at a point on the same side of the mirror as the object. Thus a real image can be projected onto a screen placed at this location. The image distance is positive, and the image is inverted, so its magnification is negative. This is a *case 1 image for mirrors*. It differs from the case 1 image for lenses only in that the image is on the same side of the mirror as the object. It is otherwise identical.



A case 1 image for a mirror. An object is farther from the converging mirror than its focal length. Rays from a common point on the object are traced using the rules in the text. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 goes through the focal point on the way toward the mirror. All three rays cross at the same point after being reflected, locating the inverted real image. Although three rays are shown, only two of the three are needed to locate the image and determine its height.

Example:

A Concave Reflector

Electric room heaters use a concave mirror to reflect infrared (IR) radiation from hot coils. Note that IR follows the same law of reflection as visible light. Given that the mirror has a radius of curvature of 50.0 cm and produces an image of the coils 3.00 m away from the mirror, where are the coils?

Strategy and Concept

We are given that the concave mirror projects a real image of the coils at an image distance $d_\mathrm{i} = 3.00\ \mathrm{m}$. The coils are the object, and we are asked to find their location—that is, to find the object distance d_o. We are also given the radius of curvature of the mirror, so that its focal length is $f = R/2 = 25.0\ \mathrm{cm}$ (positive since the mirror is concave or converging). Assuming the mirror is small compared with its radius of curvature, we can use the thin lens equations, to solve this problem.

Solution

Since d_i and f are known, thin lens equation can be used to find d_o:

Equation:

$$\frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} = \frac{1}{f}.$$

Rearranging to isolate d_o gives

Equation:

$$\frac{1}{d_\mathrm{o}} = \frac{1}{f} - \frac{1}{d_\mathrm{i}}.$$

Entering known quantities gives a value for $1/d_\mathrm{o}$:

Equation:

$$\frac{1}{d_\mathrm{o}} = \frac{1}{0.250\ \mathrm{m}} - \frac{1}{3.00\ \mathrm{m}} = \frac{3.667}{\mathrm{m}}.$$

This must be inverted to find d_o:

Equation:

$$d_\mathrm{o} = \frac{1\ \mathrm{m}}{3.667} = 27.3\ \mathrm{cm}.$$

Discussion

Note that the object (the filament) is farther from the mirror than the mirror's focal length. This is a case 1 image ($d_o > f$ and f positive), consistent with the fact that a real image is formed. You will get the most concentrated thermal energy directly in front of the mirror and 3.00 m away from it. Generally, this is not desirable, since it could cause burns. Usually, you want the rays to emerge parallel, and this is accomplished by having the filament at the focal point of the mirror.

Note that the filament here is not much farther from the mirror than its focal length and that the image produced is considerably farther away. This is exactly analogous to a slide projector. Placing a slide only slightly farther away from the projector lens than its focal length produces an image significantly farther away. As the object gets closer to the focal distance, the image gets farther away. In fact, as the object distance approaches the focal length, the image distance approaches infinity and the rays are sent out parallel to one another.

Example:

Solar Electric Generating System

One of the solar technologies used today for generating electricity is a device (called a parabolic trough or concentrating collector) that concentrates the sunlight onto a blackened pipe that contains a fluid. This heated fluid is pumped to a heat exchanger, where its heat energy is transferred to another system that is used to generate steam—and so generate electricity through a conventional steam cycle. [link] shows such a working system in southern California. Concave mirrors are used to concentrate the sunlight onto the pipe. The mirror has the approximate shape of a section of a cylinder. For the problem, assume that the mirror is exactly one-quarter of a full cylinder.

a. If we wish to place the fluid-carrying pipe 40.0 cm from the concave mirror at the mirror's focal point, what will be the radius of curvature of the mirror?
b. Per meter of pipe, what will be the amount of sunlight concentrated onto the pipe, assuming the insolation (incident solar radiation) is

$0.900\ \text{kW/m}^2$?

c. If the fluid-carrying pipe has a 2.00-cm diameter, what will be the temperature increase of the fluid per meter of pipe over a period of one minute? Assume all the solar radiation incident on the reflector is absorbed by the pipe, and that the fluid is mineral oil.

Strategy

To solve an *Integrated Concept Problem* we must first identify the physical principles involved. Part (a) is related to the current topic. Part (b) involves a little math, primarily geometry. Part (c) requires an understanding of heat and density.

Solution to (a)

To a good approximation for a concave or semi-spherical surface, the point where the parallel rays from the sun converge will be at the focal point, so $R = 2f = 80.0\ \text{cm}$.

Solution to (b)

The insolation is $900\ \text{W/m}^2$. We must find the cross-sectional area A of the concave mirror, since the power delivered is $900\ \text{W/m}^2 \times \text{A}$. The mirror in this case is a quarter-section of a cylinder, so the area for a length L of the mirror is $\text{A} = \frac{1}{4}(2\pi R)\text{L}$. The area for a length of 1.00 m is then

Equation:

$$\text{A} = \frac{\pi}{2}R(1.00\ \text{m}) = \frac{(3.14)}{2}(0.800\ \text{m})(1.00\ \text{m}) = 1.26\ \text{m}^2.$$

The insolation on the 1.00-m length of pipe is then

Equation:

$$\left(9.00 \times 10^2 \frac{\text{W}}{\text{m}^2}\right)\left(1.26\ \text{m}^2\right) = 1130\ \text{W}.$$

Solution to (c)

The increase in temperature is given by $Q = mc\,\Delta T$. The mass m of the mineral oil in the one-meter section of pipe is

Equation:

$$\begin{aligned} m &= \rho V = \rho\pi\left(\frac{d}{2}\right)^2(1.00\text{ m}) \\ &= \left(8.00\times10^2\text{ kg/m}^3\right)(3.14)(0.0100\text{ m})^2(1.00\text{ m}) \\ &= 0.251\text{ kg.} \end{aligned}$$

Therefore, the increase in temperature in one minute is

Equation:

$$\begin{aligned} \Delta T &= Q/mc \\ &= \frac{(1130\text{ W})(60.0\text{ s})}{(0.251\text{ kg})(1670\text{ J}\cdot\text{kg/}^\circ\text{C})} \\ &= 162^\circ\text{C.} \end{aligned}$$

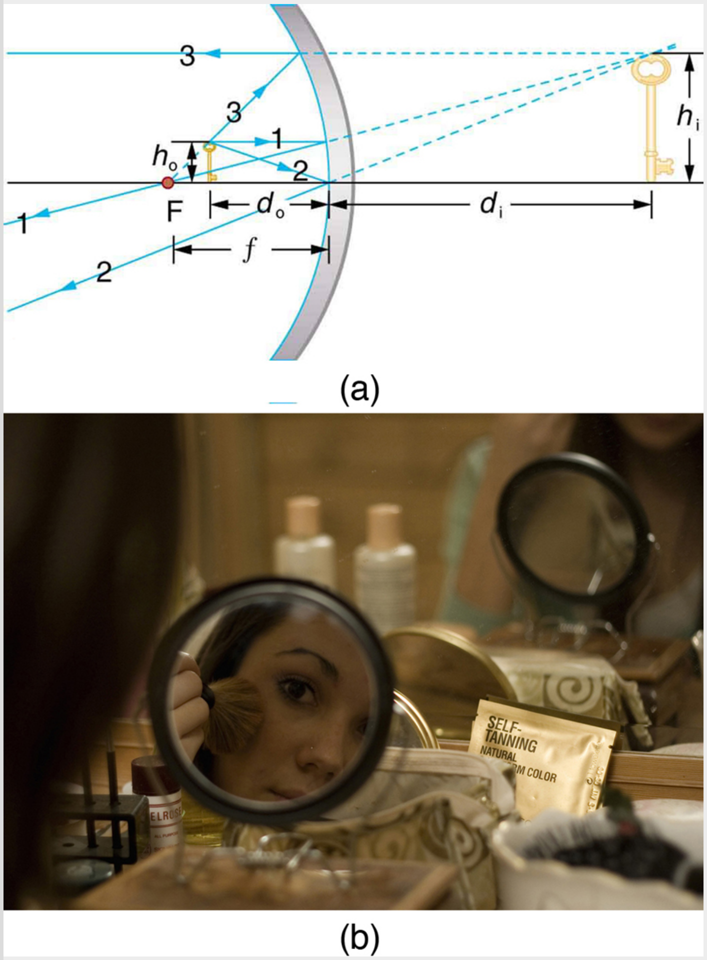
Discussion for (c)

An array of such pipes in the California desert can provide a thermal output of 250 MW on a sunny day, with fluids reaching temperatures as high as 400ºC. We are considering only one meter of pipe here, and ignoring heat losses along the pipe.

Parabolic trough collectors are used to generate electricity in southern California. (credit: kjkolb, Wikimedia Commons)

What happens if an object is closer to a concave mirror than its focal length? This is analogous to a case 2 image for lenses ($d_\text{o} < f$ and f positive), which is a magnifier. In fact, this is how makeup mirrors act as magnifiers. [link](a) uses ray tracing to locate the image of an object placed close to a concave mirror. Rays from a common point on the object

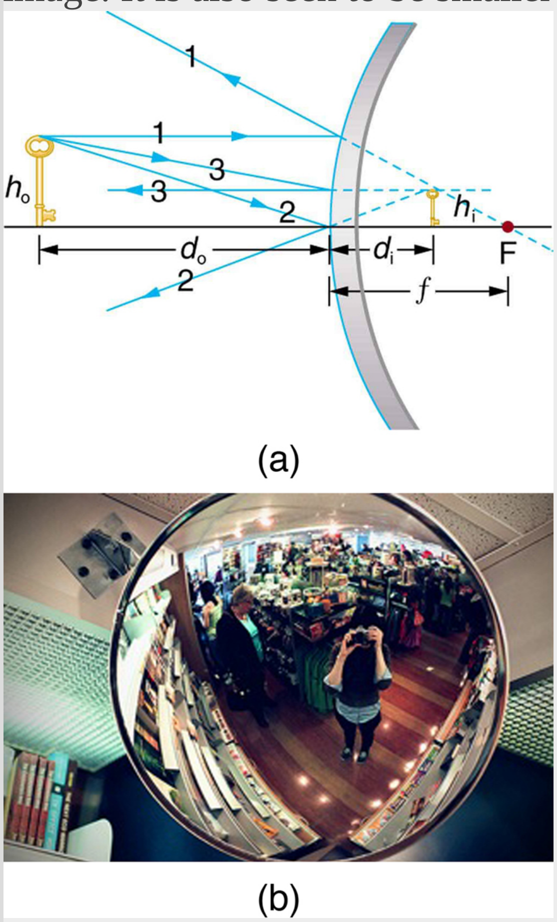
are reflected in such a manner that they appear to be coming from behind the mirror, meaning that the image is virtual and cannot be projected. As with a magnifying glass, the image is upright and larger than the object. This is a *case 2 image for mirrors* and is exactly analogous to that for lenses.



(a) Case 2 images for mirrors are formed when a converging mirror has an object closer to it than its focal length. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches the mirror as if it came from the focal point. (b) A magnifying mirror showing the reflection. (credit: Mike Melrose, Flickr)

All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be larger than the object. (b) Makeup mirrors are perhaps the most common use of a concave mirror to produce a larger, upright image.

A convex mirror is a diverging mirror (f is negative) and forms only one type of image. It is a *case 3* image—one that is upright and smaller than the object, just as for diverging lenses. [link](a) uses ray tracing to illustrate the location and size of the case 3 image for mirrors. Since the image is behind the mirror, it cannot be projected and is thus a virtual image. It is also seen to be smaller than the object.



Case 3 images for mirrors are formed by any convex mirror. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the

mirror, and ray 3 approaches toward the focal point. All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be smaller than the object. (b) Security mirrors are convex, producing a smaller, upright image. Because the image is smaller, a larger area is imaged compared to what would be observed for a flat mirror (and hence security is improved). (credit: Laura D'Alessandro, Flickr)

Example:
Image in a Convex Mirror

A keratometer is a device used to measure the curvature of the cornea, particularly for fitting contact lenses. Light is reflected from the cornea, which acts like a convex mirror, and the keratometer measures the magnification of the image. The smaller the magnification, the smaller the radius of curvature of the cornea. If the light source is 12.0 cm from the cornea and the image's magnification is 0.0320, what is the cornea's radius of curvature?

Strategy

If we can find the focal length of the convex mirror formed by the cornea, we can find its radius of curvature (the radius of curvature is twice the focal length of a spherical mirror). We are given that the object distance is

$d_o = 12.0$ cm and that $m = 0.0320$. We first solve for the image distance d_i, and then for f.
Solution
$m = -d_i/d_o$. Solving this expression for d_i gives
Equation:

$$d_i = -md_o.$$

Entering known values yields
Equation:

$$d_i = -(0.0320)(12.0 \text{ cm}) = -0.384 \text{ cm}.$$

Equation:

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_i}$$

Substituting known values,
Equation:

$$\frac{1}{f} = \frac{1}{12.0 \text{ cm}} + \frac{1}{-0.384 \text{ cm}} = \frac{-2.52}{\text{cm}}.$$

This must be inverted to find f:
Equation:

$$f = \frac{\text{cm}}{-2.52} = -0.400 \text{ cm}.$$

The radius of curvature is twice the focal length, so that
Equation:

$$R = 2 \mid f \mid = 0.800 \text{ cm}.$$

Discussion
Although the focal length f of a convex mirror is defined to be negative, we take the absolute value to give us a positive value for R. The radius of curvature found here is reasonable for a cornea. The distance from cornea

to retina in an adult eye is about 2.0 cm. In practice, many corneas are not spherical, complicating the job of fitting contact lenses. Note that the image distance here is negative, consistent with the fact that the image is behind the mirror, where it cannot be projected. In this section's Problems and Exercises, you will show that for a fixed object distance, the smaller the radius of curvature, the smaller the magnification.
The three types of images formed by mirrors (cases 1, 2, and 3) are exactly analogous to those formed by lenses, as summarized in the table at the end of Image Formation by Lenses. It is easiest to concentrate on only three types of images—then remember that concave mirrors act like convex lenses, whereas convex mirrors act like concave lenses.

Note:
Take-Home Experiment: Concave Mirrors Close to Home
Find a flashlight and identify the curved mirror used in it. Find another flashlight and shine the first flashlight onto the second one, which is turned off. Estimate the focal length of the mirror. You might try shining a flashlight on the curved mirror behind the headlight of a car, keeping the headlight switched off, and determine its focal length.

Problem-Solving Strategy for Mirrors

Step 1. Examine the situation to determine that image formation by a mirror is involved.

Step 2. Refer to the Problem-Solving Strategies for Lenses. The same strategies are valid for mirrors as for lenses with one qualification—use the ray tracing rules for mirrors listed earlier in this section.

Section Summary

- The characteristics of an image formed by a flat mirror are: (a) The image and object are the same distance from the mirror, (b) The image

is a virtual image, and (c) The image is situated behind the mirror.
- Image length is half the radius of curvature.
  Equation:

$$f = \frac{R}{2}$$

- A convex mirror is a diverging mirror and forms only one type of image, namely a virtual image.

Conceptual Questions

Exercise:

Problem:

What are the differences between real and virtual images? How can you tell (by looking) whether an image formed by a single lens or mirror is real or virtual?

Exercise:

Problem:

Can you see a virtual image? Can you photograph one? Can one be projected onto a screen with additional lenses or mirrors? Explain your responses.

Exercise:

Problem:

Is it necessary to project a real image onto a screen for it to exist?

Exercise:

Problem:

At what distance is an image *always* located—at d_o, d_i, or f?

Exercise:

Problem:

Under what circumstances will an image be located at the focal point of a lens or mirror?

Exercise:

Problem:

What is meant by a negative magnification? What is meant by a magnification that is less than 1 in magnitude?

Exercise:

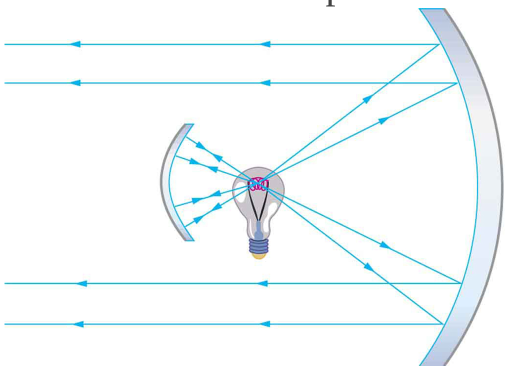
Problem:

Can a case 1 image be larger than the object even though its magnification is always negative? Explain.

Exercise:

Problem:

[link] shows a light bulb between two mirrors. One mirror produces a beam of light with parallel rays; the other keeps light from escaping without being put into the beam. Where is the filament of the light in relation to the focal point or radius of curvature of each mirror?



The two mirrors trap most of the bulb's light and form a directional beam as in a headlight.

Exercise:

Problem:

Devise an arrangement of mirrors allowing you to see the back of your head. What is the minimum number of mirrors needed for this task?

Exercise:

Problem:

If you wish to see your entire body in a flat mirror (from head to toe), how tall should the mirror be? Does its size depend upon your distance away from the mirror? Provide a sketch.

Exercise:

Problem:

It can be argued that a flat mirror has an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?

Exercise:

Problem:

Why are diverging mirrors often used for rear-view mirrors in vehicles? What is the main disadvantage of using such a mirror compared with a flat one?

Problems & Exercises

Exercise:

Problem:

What is the focal length of a makeup mirror that has a power of 1.50 D?

Solution:

+0.667 m

Exercise:

Problem:

Some telephoto cameras use a mirror rather than a lens. What radius of curvature mirror is needed to replace a 800 mm focal length telephoto lens?

Exercise:

Problem:

(a) Calculate the focal length of the mirror formed by the shiny back of a spoon that has a 3.00 cm radius of curvature. (b) What is its power in diopters?

Solution:

(a) -1.5×10^{-2} m

(b)-66.7 D

Exercise:

Problem:

Find the magnification of the heater element in [link]. Note that its large magnitude helps spread out the reflected energy.

Exercise:

Problem:

What is the focal length of a makeup mirror that produces a magnification of 1.50 when a person's face is 12.0 cm away? Explicitly show how you follow the steps in the Problem-Solving Strategy for Mirrors.

Solution:

+0.360 m (concave)

Exercise:

Problem:

A shopper standing 3.00 m from a convex security mirror sees his image with a magnification of 0.250. (a) Where is his image? (b) What is the focal length of the mirror? (c) What is its radius of curvature? Explicitly show how you follow the steps in the Problem-Solving Strategy for Mirrors.

Exercise:

Problem:

An object 1.50 cm high is held 3.00 cm from a person's cornea, and its reflected image is measured to be 0.167 cm high. (a) What is the magnification? (b) Where is the image? (c) Find the radius of curvature of the convex mirror formed by the cornea. (Note that this technique is used by optometrists to measure the curvature of the cornea for contact lens fitting. The instrument used is called a keratometer, or curve measurer.)

Solution:

(a) +0.111

(b) -0.334 cm (behind "mirror")

(c) 0.752cm

Exercise:

Problem:

Ray tracing for a flat mirror shows that the image is located a distance behind the mirror equal to the distance of the object from the mirror. This is stated $d_{\text{i}} = -d_{\text{o}}$, since this is a negative image distance (it is a virtual image). (a) What is the focal length of a flat mirror? (b) What is its power?

Exercise:

Problem:

Show that for a flat mirror $h_{\mathrm{i}} = h_{\mathrm{o}}$, knowing that the image is a distance behind the mirror equal in magnitude to the distance of the object from the mirror.

Solution:
Equation:

$$m = \frac{h_{\mathrm{i}}}{h_{\mathrm{o}}} = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = -\frac{-d_{\mathrm{o}}}{d_{\mathrm{o}}} = \frac{d_{\mathrm{o}}}{d_{\mathrm{o}}} = 1 \Rightarrow h_{\mathrm{i}} = h_{\mathrm{o}}$$

Exercise:

Problem:

Use the law of reflection to prove that the focal length of a mirror is half its radius of curvature. That is, prove that $f = R/2$. Note this is true for a spherical mirror only if its diameter is small compared with its radius of curvature.

Exercise:

Problem:

Referring to the electric room heater considered in the first example in this section, calculate the intensity of IR radiation in $\mathrm{W/m^2}$ projected by the concave mirror on a person 3.00 m away. Assume that the heating element radiates 1500 W and has an area of $100\ \mathrm{cm}^2$, and that half of the radiated power is reflected and focused by the mirror.

Solution:

$6.82\ \mathrm{kW/m^2}$

Exercise:

Problem:

Consider a 250-W heat lamp fixed to the ceiling in a bathroom. If the filament in one light burns out then the remaining three still work. Construct a problem in which you determine the resistance of each filament in order to obtain a certain intensity projected on the bathroom floor. The ceiling is 3.0 m high. The problem will need to involve concave mirrors behind the filaments. Your instructor may wish to guide you on the level of complexity to consider in the electrical components.

Glossary

converging mirror
: a concave mirror in which light rays that strike it parallel to its axis converge at one or more points along the axis

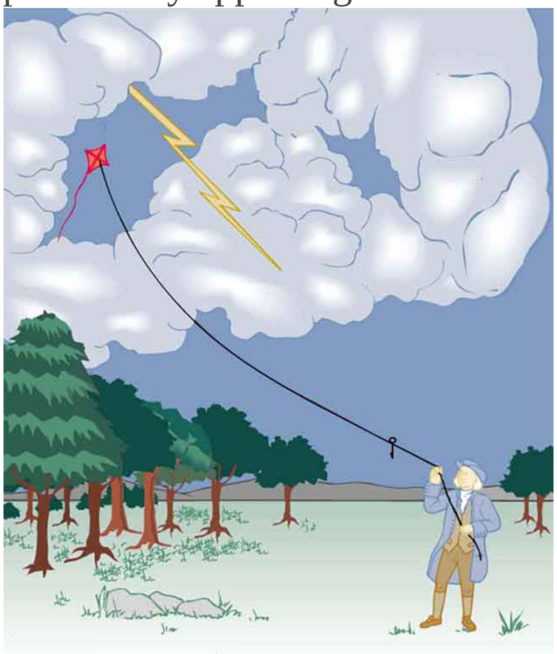
diverging mirror
: a convex mirror in which light rays that strike it parallel to its axis bend away (diverge) from its axis

law of reflection
: angle of reflection equals the angle of incidence

Introduction to Electric Charge and Electric Field class="introduction"

Static electricity from this plastic slide causes the child’s hair to stand on end. The sliding motion stripped electrons away from the child’s body, leaving an excess of positive charges, which repel each other along each strand of hair. (credit: Ken Bosma/Wikimedia Commons)

The image of American politician and scientist Benjamin Franklin (1706–1790) flying a kite in a thunderstorm is familiar to every schoolchild. (See [link].) In this experiment, Franklin demonstrated a connection between lightning and **static electricity**. Sparks were drawn from a key hung on a kite string during an electrical storm. These sparks were like those produced by static electricity, such as the spark that jumps from your finger to a metal doorknob after you walk across a wool carpet. What Franklin demonstrated in his dangerous experiment was a connection between phenomena on two different scales: one the grand power of an electrical storm, the other an effect of more human proportions. Connections like this one reveal the underlying unity of the laws of nature, an aspect we humans find particularly appealing.



When Benjamin Franklin demonstrated that lightning was related to static electricity, he made a connection that is now part of the evidence that all directly experienced forces except the gravitational force are manifestations of the electromagnetic force.

Much has been written about Franklin. His experiments were only part of the life of a man who was a scientist, inventor, revolutionary, statesman, and writer. Franklin's experiments were not performed in isolation, nor were they the only ones to reveal connections.

For example, the Italian scientist Luigi Galvani (1737–1798) performed a series of experiments in which static electricity was used to stimulate contractions of leg muscles of dead frogs, an effect already known in humans subjected to static discharges. But Galvani also found that if he joined two metal wires (say copper and zinc) end to end and touched the other ends to muscles, he produced the same effect in frogs as static discharge. Alessandro Volta (1745–1827), partly inspired by Galvani's work, experimented with various combinations of metals and developed the battery.

During the same era, other scientists made progress in discovering fundamental connections. The periodic table was developed as the systematic properties of the elements were discovered. This influenced the development and refinement of the concept of atoms as the basis of matter. Such submicroscopic descriptions of matter also help explain a great deal more.

Atomic and molecular interactions, such as the forces of friction, cohesion, and adhesion, are now known to be manifestations of the **electromagnetic force**. Static electricity is just one aspect of the electromagnetic force, which also includes moving electricity and magnetism.

All the macroscopic forces that we experience directly, such as the sensations of touch and the tension in a rope, are due to the electromagnetic force, one of the four fundamental forces in nature. The gravitational force, another fundamental force, is actually sensed through the electromagnetic interaction of molecules, such as between those in our feet and those on the top of a bathroom scale. (The other two fundamental forces, the strong nuclear force and the weak nuclear force, cannot be sensed on the human scale.)

This chapter begins the study of electromagnetic phenomena at a fundamental level. The next several chapters will cover static electricity, moving electricity, and magnetism—collectively known as electromagnetism. In this chapter, we begin with the study of electric phenomena due to charges that are at least temporarily stationary, called electrostatics, or static electricity.

Glossary

static electricity
: a buildup of electric charge on the surface of an object

electromagnetic force
: one of the four fundamental forces of nature; the electromagnetic force consists of static electricity, moving electricity and magnetism

Static Electricity and Charge: Conservation of Charge

- Define electric charge, and describe how the two types of charge interact.
- Describe three common situations that generate static electricity.
- State the law of conservation of charge.

Borneo amber was mined in Sabah, Malaysia, from shale-sandstone-mudstone veins. When a piece of amber is rubbed with a piece of silk, the amber gains more electrons, giving it a net negative charge. At the same time, the silk, having lost electrons, becomes positively charged. (credit: Sebakoamber, Wikimedia Commons)

What makes plastic wrap cling? Static electricity. Not only are applications of static electricity common these days, its existence has been known since ancient times. The first record of its effects dates to ancient Greeks who noted more than 500 years B.C. that polishing amber temporarily enabled it

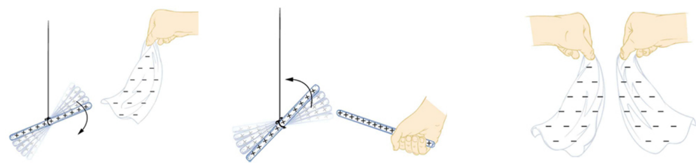
to attract bits of straw (see [link]). The very word *electric* derives from the Greek word for amber (*electron*).

Many of the characteristics of static electricity can be explored by rubbing things together. Rubbing creates the spark you get from walking across a wool carpet, for example. Static cling generated in a clothes dryer and the attraction of straw to recently polished amber also result from rubbing. Similarly, lightning results from air movements under certain weather conditions. You can also rub a balloon on your hair, and the static electricity created can then make the balloon cling to a wall. We also have to be cautious of static electricity, especially in dry climates. When we pump gasoline, we are warned to discharge ourselves (after sliding across the seat) on a metal surface before grabbing the gas nozzle. Attendants in hospital operating rooms must wear booties with aluminum foil on the bottoms to avoid creating sparks which may ignite the oxygen being used.

Some of the most basic characteristics of static electricity include:

- The effects of static electricity are explained by a physical quantity not previously introduced, called electric charge.
- There are only two types of charge, one called positive and the other called negative.
- Like charges repel, whereas unlike charges attract.
- The force between charges decreases with distance.

How do we know there are two types of **electric charge**? When various materials are rubbed together in controlled ways, certain combinations of materials always produce one type of charge on one material and the opposite type on the other. By convention, we call one type of charge "positive", and the other type "negative." For example, when glass is rubbed with silk, the glass becomes positively charged and the silk negatively charged. Since the glass and silk have opposite charges, they attract one another like clothes that have rubbed together in a dryer. Two glass rods rubbed with silk in this manner will repel one another, since each rod has positive charge on it. Similarly, two silk cloths so rubbed will repel, since both cloths have negative charge. [link] shows how these simple materials can be used to explore the nature of the force between charges.



A glass rod becomes positively charged when rubbed with silk, while the silk becomes negatively charged. (a) The glass rod is attracted to the silk because their charges are opposite. (b) Two similarly charged glass rods repel. (c) Two similarly charged silk cloths repel.

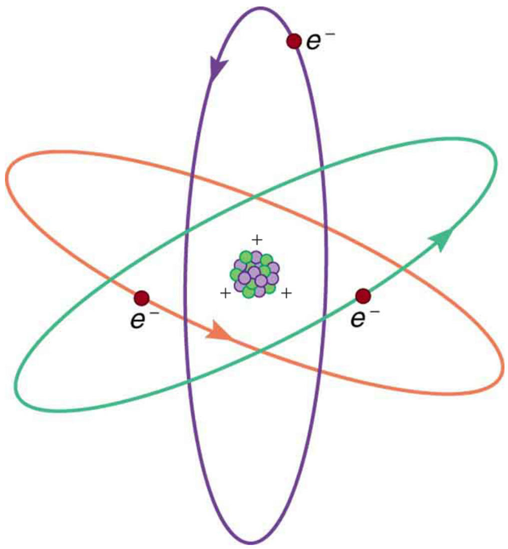
More sophisticated questions arise. Where do these charges come from? Can you create or destroy charge? Is there a smallest unit of charge? Exactly how does the force depend on the amount of charge and the distance between charges? Such questions obviously occurred to Benjamin Franklin and other early researchers, and they interest us even today.

Charge Carried by Electrons and Protons

Franklin wrote in his letters and books that he could see the effects of electric charge but did not understand what caused the phenomenon. Today we have the advantage of knowing that normal matter is made of atoms, and that atoms contain positive and negative charges, usually in equal amounts.

[link] shows a simple model of an atom with negative **electrons** orbiting its positive nucleus. The nucleus is positive due to the presence of positively charged **protons**. Nearly all charge in nature is due to electrons and protons, which are two of the three building blocks of most matter. (The third is the neutron, which is neutral, carrying no charge.) Other charge-carrying particles are observed in cosmic rays and nuclear decay, and are created in

particle accelerators. All but the electron and proton survive only a short time and are quite rare by comparison.



This simplified (and not to scale) view of an atom is called the planetary model of the atom. Negative electrons orbit a much heavier positive nucleus, as the planets orbit the much heavier sun. There the similarity ends, because forces in the atom are electromagnetic, whereas those in the planetary system are gravitational. Normal macroscopic amounts of matter contain immense numbers of atoms and molecules and, hence, even greater numbers of individual

negative and positive charges.

The charges of electrons and protons are identical in magnitude but opposite in sign. Furthermore, all charged objects in nature are integral multiples of this basic quantity of charge, meaning that all charges are made of combinations of a basic unit of charge. Usually, charges are formed by combinations of electrons and protons. The magnitude of this basic charge is

Equation:

$$\mid q_e \mid = 1.60 \times 10^{-19}\ \text{C}.$$

The symbol q is commonly used for charge and the subscript e indicates the charge of a single electron (or proton).

The SI unit of charge is the coulomb (C). The number of protons needed to make a charge of 1.00 C is

Equation:

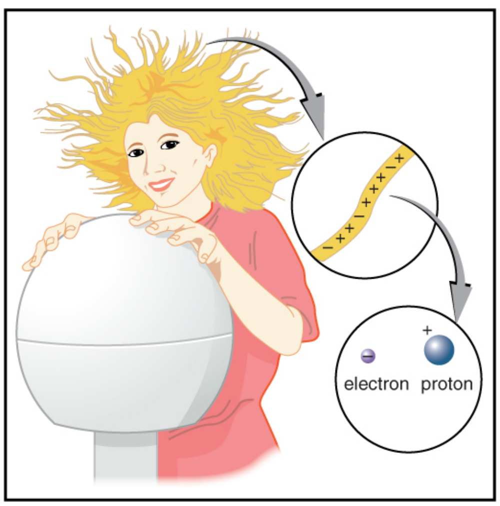
$$1.00\ \text{C} \times \frac{1\ \text{proton}}{1.60 \times 10^{-19}\ \text{C}} = 6.25 \times 10^{18}\ \text{protons}.$$

Similarly, 6.25×10^{18} electrons have a combined charge of −1.00 coulomb. Just as there is a smallest bit of an element (an atom), there is a smallest bit of charge. There is no directly observed charge smaller than $\mid q_e \mid$ (see Things Great and Small: The Submicroscopic Origin of Charge), and all observed charges are integral multiples of $\mid q_e \mid$.

Note:
Things Great and Small: The Submicroscopic Origin of Charge

With the exception of exotic, short-lived particles, all charge in nature is carried by electrons and protons. Electrons carry the charge we have named negative. Protons carry an equal-magnitude charge that we call positive. (See [link].) Electron and proton charges are considered fundamental building blocks, since all other charges are integral multiples of those carried by electrons and protons. Electrons and protons are also two of the three fundamental building blocks of ordinary matter. The neutron is the third and has zero total charge.

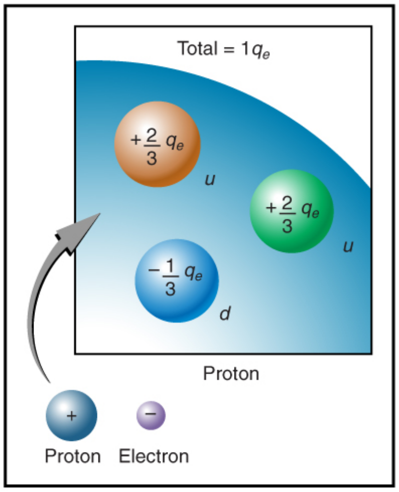
[link] shows a person touching a Van de Graaff generator and receiving excess positive charge. The expanded view of a hair shows the existence of both types of charges but an excess of positive. The repulsion of these positive like charges causes the strands of hair to repel other strands of hair and to stand up. The further blowup shows an artist's conception of an electron and a proton perhaps found in an atom in a strand of hair.



When this person touches a Van de Graaff generator, she receives an excess of positive charge, causing her hair to stand on end. The charges in

one hair are shown. An artist’s conception of an electron and a proton illustrate the particles carrying the negative and positive charges. We cannot really see these particles with visible light because they are so small (the electron seems to be an infinitesimal point), but we know a great deal about their measurable properties, such as the charges they carry.

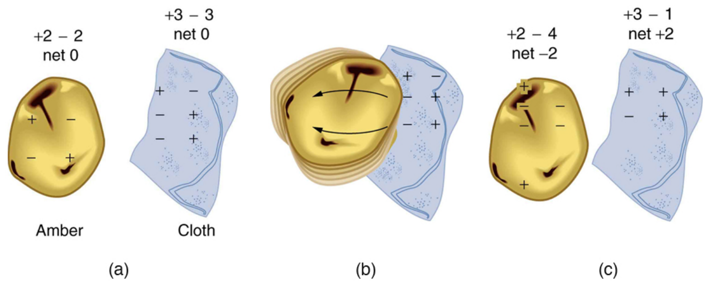
The electron seems to have no substructure; in contrast, when the substructure of protons is explored by scattering extremely energetic electrons from them, it appears that there are point-like particles inside the proton. These sub-particles, named quarks, have never been directly observed, but they are believed to carry fractional charges as seen in [link]. Charges on electrons and protons and all other directly observable particles are unitary, but these quark substructures carry charges of either $-\frac{1}{3}$ or $+\frac{2}{3}$. There are continuing attempts to observe fractional charge directly and to learn of the properties of quarks, which are perhaps the ultimate substructure of matter.



Artist's conception of fractional quark charges inside a proton. A group of three quark charges add up to the single positive charge on the proton: $-\frac{1}{3}q_e + \frac{2}{3}q_e + \frac{2}{3}q_e = +1q_e$.

Separation of Charge in Atoms

Charges in atoms and molecules can be separated—for example, by rubbing materials together. Some atoms and molecules have a greater affinity for electrons than others and will become negatively charged by close contact in rubbing, leaving the other material positively charged. (See [link].) Positive charge can similarly be induced by rubbing. Methods other than rubbing can also separate charges. Batteries, for example, use combinations of substances that interact in such a way as to separate charges. Chemical interactions may transfer negative charge from one substance to the other, making one battery terminal negative and leaving the first one positive.



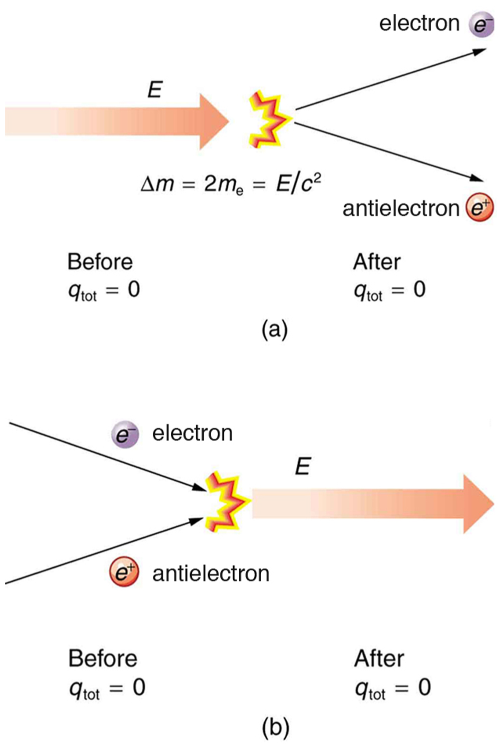
When materials are rubbed together, charges can be separated, particularly if one material has a greater affinity for electrons than another. (a) Both the amber and cloth are originally neutral, with equal positive and negative charges. Only a tiny fraction of the charges are involved, and only a few of them are shown here. (b) When rubbed together, some negative charge is transferred to the amber, leaving the cloth with a net positive charge. (c) When separated, the amber and cloth now have net charges, but the absolute value of the net positive and negative charges will be equal.

No charge is actually created or destroyed when charges are separated as we have been discussing. Rather, existing charges are moved about. In fact, in all situations the total amount of charge is always constant. This universally obeyed law of nature is called the **law of conservation of charge**.

Note:
Law of Conservation of Charge
Total charge is constant in any process.

In more exotic situations, such as in particle accelerators, mass, Δm, can be created from energy in the amount $\Delta m = \frac{E}{c^2}$. Sometimes, the created mass is charged, such as when an electron is created. Whenever a charged particle is created, another having an opposite charge is always created along with it, so that the total charge created is zero. Usually, the two particles are "matter-antimatter" counterparts. For example, an antielectron would usually be created at the same time as an electron. The antielectron has a positive charge (it is called a positron), and so the total charge created is zero. (See [link].) All particles have antimatter counterparts with opposite signs. When matter and antimatter counterparts are brought together, they completely annihilate one another. By annihilate, we mean that the mass of the two particles is converted to energy E, again obeying the relationship $\Delta m = \frac{E}{c^2}$. Since the two particles have equal and opposite charge, the total charge is zero before and after the annihilation; thus, total charge is conserved.

Note:
Making Connections: Conservation Laws
Only a limited number of physical quantities are universally conserved. Charge is one—energy, momentum, and angular momentum are others. Because they are conserved, these physical quantities are used to explain more phenomena and form more connections than other, less basic quantities. We find that conserved quantities give us great insight into the rules followed by nature and hints to the organization of nature. Discoveries of conservation laws have led to further discoveries, such as the weak nuclear force and the quark substructure of protons and other particles.



(a) When enough energy is present, it can be converted into matter. Here the matter created is an electron–antielectron pair. (m_e is the electron's mass.) The total charge before and after this event is zero. (b) When matter and antimatter collide, they annihilate each other; the total charge is conserved at zero before and after the annihilation.

The law of conservation of charge is absolute—it has never been observed to be violated. Charge, then, is a special physical quantity, joining a very

short list of other quantities in nature that are always conserved. Other conserved quantities include energy, momentum, and angular momentum.

Note:
PhET Explorations: Balloons and Static Electricity
Why does a balloon stick to your sweater? Rub a balloon on a sweater, then let go of the balloon and it flies over and sticks to the sweater. View the charges in the sweater, balloons, and the wall.
https://phet.colorado.edu/sims/html/balloons-and-static-electricity/latest/balloons-and-static-electricity_en.html

Section Summary

- There are only two types of charge, which we call positive and negative.
- Like charges repel, unlike charges attract, and the force between charges decreases with the square of the distance.
- The vast majority of positive charge in nature is carried by protons, while the vast majority of negative charge is carried by electrons.
- The electric charge of one electron is equal in magnitude and opposite in sign to the charge of one proton.
- An ion is an atom or molecule that has nonzero total charge due to having unequal numbers of electrons and protons.
- The SI unit for charge is the coulomb (C), with protons and electrons having charges of opposite sign but equal magnitude; the magnitude of this basic charge $\mid q_e \mid$ is
  Equation:

  $$\mid q_e \mid = 1.60 \times 10^{-19}\text{ C}.$$

- Whenever charge is created or destroyed, equal amounts of positive and negative are involved.
- Most often, existing charges are separated from neutral objects to obtain some net charge.

- Both positive and negative charges exist in neutral objects and can be separated by rubbing one object with another. For macroscopic objects, negatively charged means an excess of electrons and positively charged means a depletion of electrons.
- The law of conservation of charge ensures that whenever a charge is created, an equal charge of the opposite sign is created at the same time.

Conceptual Questions

Exercise:

Problem:

There are very large numbers of charged particles in most objects. Why, then, don't most objects exhibit static electricity?

Exercise:

Problem:

Why do most objects tend to contain nearly equal numbers of positive and negative charges?

Problems & Exercises

Exercise:

Problem:

Common static electricity involves charges ranging from nanocoulombs to microcoulombs. (a) How many electrons are needed to form a charge of -2.00 nC (b) How many electrons must be removed from a neutral object to leave a net charge of $0.500 \ \mu\text{C}$?

Solution:

(a) 1.25×10^{10}

(b) 3.13×10^{12}

Exercise:

Problem:

If 1.80×10^{20} electrons move through a pocket calculator during a full day's operation, how many coulombs of charge moved through it?

Exercise:

Problem:

To start a car engine, the car battery moves 3.75×10^{21} electrons through the starter motor. How many coulombs of charge were moved?

Solution:

-600 C

Exercise:

Problem:

A certain lightning bolt moves 40.0 C of charge. How many fundamental units of charge $|q_e|$ is this?

Glossary

electric charge
: a physical property of an object that causes it to be attracted toward or repelled from another charged object; each charged object generates and is influenced by a force called an electromagnetic force

law of conservation of charge
: states that whenever a charge is created, an equal amount of charge with the opposite sign is created simultaneously

electron

a particle orbiting the nucleus of an atom and carrying the smallest unit of negative charge

proton

a particle in the nucleus of an atom and carrying a positive charge equal in magnitude and opposite in sign to the amount of negative charge carried by an electron

Conductors and Insulators

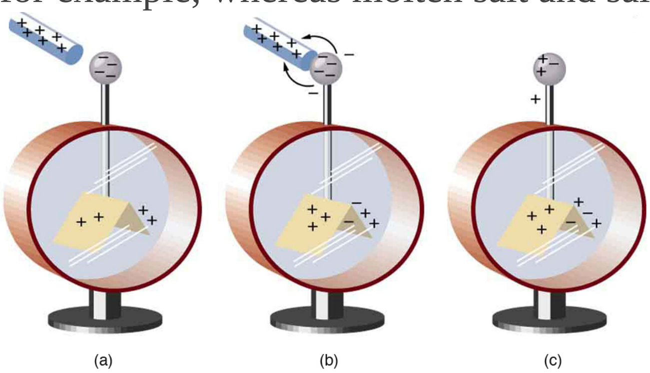
- Define conductor and insulator, explain the difference, and give examples of each.
- Describe three methods for charging an object.
- Explain what happens to an electric force as you move farther from the source.
- Define polarization.

This power adapter uses metal wires and connectors to conduct electricity from the wall socket to a laptop computer. The conducting wires allow electrons to move freely through the cables, which are shielded by rubber and plastic. These materials act as insulators that don’t allow electric charge to escape outward. (credit: Evan-Amos, Wikimedia Commons)

Some substances, such as metals and salty water, allow charges to move through them with relative ease. Some of the electrons in metals and similar conductors are not bound to individual atoms or sites in the material. These **free electrons** can move through the material much as air moves through loose sand. Any substance that has free electrons and allows charge to move

relatively freely through it is called a **conductor**. The moving electrons may collide with fixed atoms and molecules, losing some energy, but they can move in a conductor. Superconductors allow the movement of charge without any loss of energy. Salty water and other similar conducting materials contain free ions that can move through them. An ion is an atom or molecule having a positive or negative (nonzero) total charge. In other words, the total number of electrons is not equal to the total number of protons.

Other substances, such as glass, do not allow charges to move through them. These are called **insulators**. Electrons and ions in insulators are bound in the structure and cannot move easily—as much as 10^{23} times more slowly than in conductors. Pure water and dry table salt are insulators, for example, whereas molten salt and salty water are conductors.



An electroscope is a favorite instrument in physics demonstrations and student laboratories. It is typically made with gold foil leaves hung from a (conducting) metal stem and is insulated from the room air in a glass-walled container. (a) A positively charged glass rod is brought near the tip of the electroscope, attracting electrons to the top and leaving a net positive charge on the leaves. Like charges in the light flexible gold leaves

repel, separating them. (b) When the rod is touched against the ball, electrons are attracted and transferred, reducing the net charge on the glass rod but leaving the electroscope positively charged. (c) The excess charges are evenly distributed in the stem and leaves of the electroscope once the glass rod is removed.

Charging by Contact

[link] shows an electroscope being charged by touching it with a positively charged glass rod. Because the glass rod is an insulator, it must actually touch the electroscope to transfer charge to or from it. (Note that the extra positive charges reside on the surface of the glass rod as a result of rubbing it with silk before starting the experiment.) Since only electrons move in metals, we see that they are attracted to the top of the electroscope. There, some are transferred to the positive rod by touch, leaving the electroscope with a net positive charge.

Electrostatic repulsion in the leaves of the charged electroscope separates them. The electrostatic force has a horizontal component that results in the leaves moving apart as well as a vertical component that is balanced by the gravitational force. Similarly, the electroscope can be negatively charged by contact with a negatively charged object.

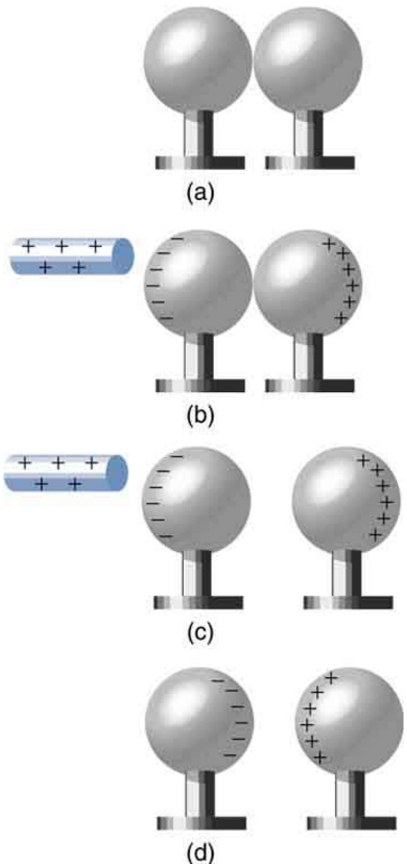
Charging by Induction

It is not necessary to transfer excess charge directly to an object in order to charge it. [link] shows a method of **induction** wherein a charge is created in a nearby object, without direct contact. Here we see two neutral metal spheres in contact with one another but insulated from the rest of the world.

A positively charged rod is brought near one of them, attracting negative charge to that side, leaving the other sphere positively charged.

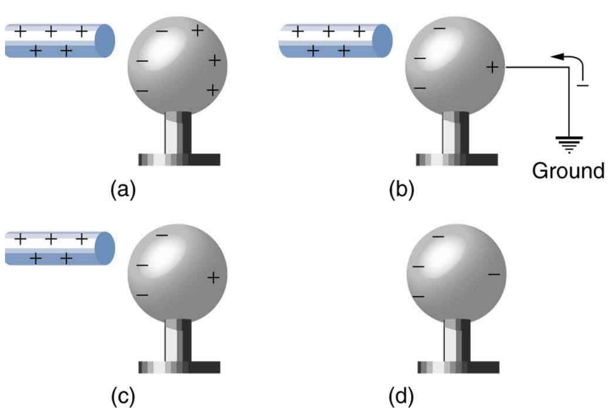
This is an example of induced **polarization** of neutral objects. Polarization is the separation of charges in an object that remains neutral. If the spheres are now separated (before the rod is pulled away), each sphere will have a net charge. Note that the object closest to the charged rod receives an opposite charge when charged by induction. Note also that no charge is removed from the charged rod, so that this process can be repeated without depleting the supply of excess charge.

Another method of charging by induction is shown in [link]. The neutral metal sphere is polarized when a charged rod is brought near it. The sphere is then grounded, meaning that a conducting wire is run from the sphere to the ground. Since the earth is large and most ground is a good conductor, it can supply or accept excess charge easily. In this case, electrons are attracted to the sphere through a wire called the ground wire, because it supplies a conducting path to the ground. The ground connection is broken before the charged rod is removed, leaving the sphere with an excess charge opposite to that of the rod. Again, an opposite charge is achieved when charging by induction and the charged rod loses none of its excess charge.



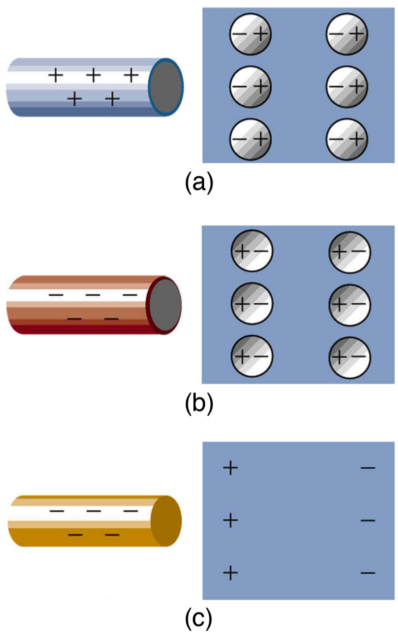
Charging by induction. (a) Two uncharged or neutral metal spheres are in contact with each other but insulated from the rest of the world. (b) A positively charged glass rod is brought near the sphere on the left, attracting negative charge and leaving the other sphere positively charged. (c) The

spheres are separated before the rod is removed, thus separating negative and positive charge. (d) The spheres retain net charges after the inducing rod is removed—without ever having been touched by a charged object.



Charging by induction, using a ground connection. (a) A positively charged rod is brought near a neutral metal sphere, polarizing it. (b) The sphere is grounded, allowing electrons to be attracted from the earth's ample supply. (c) The ground connection is broken. (d) The positive rod is

removed, leaving the sphere with an induced negative charge.



Both positive and negative objects attract a neutral object by polarizing its molecules. (a) A positive object brought near a neutral insulator polarizes its molecules. There is a slight shift in the distribution of the electrons orbiting the molecule, with

unlike charges being brought nearer and like charges moved away. Since the electrostatic force decreases with distance, there is a net attraction. (b) A negative object produces the opposite polarization, but again attracts the neutral object. (c) The same effect occurs for a conductor; since the unlike charges are closer, there is a net attraction.

Neutral objects can be attracted to any charged object. The pieces of straw attracted to polished amber are neutral, for example. If you run a plastic comb through your hair, the charged comb can pick up neutral pieces of paper. [link] shows how the polarization of atoms and molecules in neutral objects results in their attraction to a charged object.

When a charged rod is brought near a neutral substance, an insulator in this case, the distribution of charge in atoms and molecules is shifted slightly. Opposite charge is attracted nearer the external charged rod, while like charge is repelled. Since the electrostatic force decreases with distance, the repulsion of like charges is weaker than the attraction of unlike charges, and so there is a net attraction. Thus a positively charged glass rod attracts neutral pieces of paper, as will a negatively charged rubber rod. Some

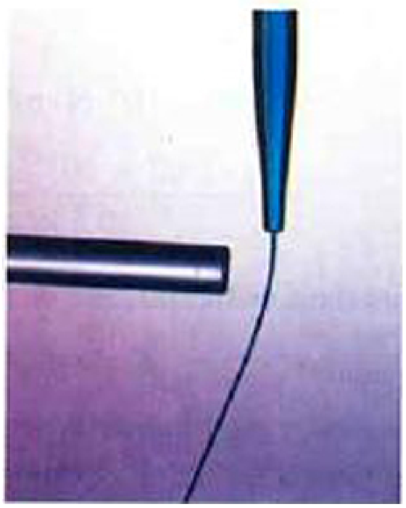
molecules, like water, are polar molecules. Polar molecules have a natural or inherent separation of charge, although they are neutral overall. Polar molecules are particularly affected by other charged objects and show greater polarization effects than molecules with naturally uniform charge distributions.

Exercise:

Check Your Understanding

Problem:

Can you explain the attraction of water to the charged rod in the figure below?



Solution:

Answer

Water molecules are polarized, giving them slightly positive and slightly negative sides. This makes water even more susceptible to a charged rod's attraction. As the water flows downward, due to the force of gravity, the charged conductor exerts a net attraction to the opposite charges in the stream of water, pulling it closer.

Note:

PhET Explorations: John Travoltage

Make sparks fly with John Travoltage. Wiggle Johnnie's foot and he picks up charges from the carpet. Bring his hand close to the door knob and get rid of the excess charge.
https://phet.colorado.edu/sims/html/john-travoltage/latest/john-travoltage_en.html

Section Summary

- Polarization is the separation of positive and negative charges in a neutral object.
- A conductor is a substance that allows charge to flow freely through its atomic structure.
- An insulator holds charge within its atomic structure.
- Objects with like charges repel each other, while those with unlike charges attract each other.
- A conducting object is said to be grounded if it is connected to the Earth through a conductor. Grounding allows transfer of charge to and from the earth's large reservoir.
- Objects can be charged by contact with another charged object and obtain the same sign charge.
- If an object is temporarily grounded, it can be charged by induction, and obtains the opposite sign charge.
- Polarized objects have their positive and negative charges concentrated in different areas, giving them a non-symmetrical charge.
- Polar molecules have an inherent separation of charge.

Conceptual Questions

Exercise:

Problem:

An eccentric inventor attempts to levitate by first placing a large negative charge on himself and then putting a large positive charge on the ceiling of his workshop. Instead, while attempting to place a large negative charge on himself, his clothes fly off. Explain.

Exercise:

Problem:

If you have charged an electroscope by contact with a positively charged object, describe how you could use it to determine the charge of other objects. Specifically, what would the leaves of the electroscope do if other charged objects were brought near its knob?

Exercise:

Problem:

When a glass rod is rubbed with silk, it becomes positive and the silk becomes negative—yet both attract dust. Does the dust have a third type of charge that is attracted to both positive and negative? Explain.

Exercise:

Problem:

Why does a car always attract dust right after it is polished? (Note that car wax and car tires are insulators.)

Exercise:

Problem:

Describe how a positively charged object can be used to give another object a negative charge. What is the name of this process?

Exercise:

Problem:

What is grounding? What effect does it have on a charged conductor? On a charged insulator?

Problems & Exercises

Exercise:

Problem:

Suppose a speck of dust in an electrostatic precipitator has 1.0000×10^{12} protons in it and has a net charge of –5.00 nC (a very large charge for a small speck). How many electrons does it have?

Solution:

1.03×10^{12}

Exercise:

Problem:

An amoeba has 1.00×10^{16} protons and a net charge of 0.300 pC. (a) How many fewer electrons are there than protons? (b) If you paired them up, what fraction of the protons would have no electrons?

Exercise:

Problem:

A 50.0 g ball of copper has a net charge of $2.00\ \mu\text{C}$. What fraction of the copper's electrons has been removed? (Each copper atom has 29 protons, and copper has an atomic mass of 63.5.)

Solution:

9.09×10^{-13}

Exercise:

Problem:

What net charge would you place on a 100 g piece of sulfur if you put an extra electron on 1 in 10^{12} of its atoms? (Sulfur has an atomic mass of 32.1.)

Exercise:

Problem:

How many coulombs of positive charge are there in 4.00 kg of plutonium, given its atomic mass is 244 and that each plutonium atom has 94 protons?

Solution:

1.48×10^{8} C

Glossary

free electron
: an electron that is free to move away from its atomic orbit

conductor
: a material that allows electrons to move separately from their atomic orbits

insulator
: a material that holds electrons securely within their atomic orbits

grounded
: when a conductor is connected to the Earth, allowing charge to freely flow to and from Earth's unlimited reservoir

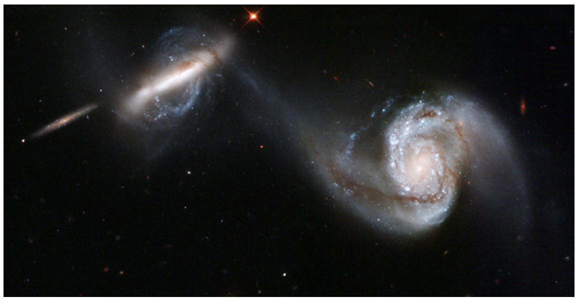
induction
: the process by which an electrically charged object brought near a neutral object creates a charge in that object

polarization
: slight shifting of positive and negative charges to opposite sides of an atom or molecule

electrostatic repulsion
: the phenomenon of two objects with like charges repelling each other

Coulomb’s Law

- State Coulomb’s law in terms of how the electrostatic force changes with the distance between two objects.
- Calculate the electrostatic force between two charged point forces, such as electrons or protons.
- Compare the electrostatic force to the gravitational attraction for a proton and an electron; for a human and the Earth.



This NASA image of Arp 87 shows the result of a strong gravitational attraction between two galaxies. In contrast, at the subatomic level, the electrostatic attraction between two objects, such as an electron and a proton, is far greater than their mutual attraction due to gravity. (credit: NASA/HST)

Through the work of scientists in the late 18th century, the main features of the **electrostatic force**—the existence of two types of charge, the observation that like charges repel, unlike charges attract, and the decrease of force with distance—were eventually refined, and expressed as a mathematical formula. The mathematical formula for the electrostatic force is called **Coulomb’s law** after the French physicist Charles Coulomb (1736–1806), who performed experiments and first proposed a formula to calculate it.

Note:
Coulomb’s Law
Equation:

$$F = k\frac{|q_1 q_2|}{r^2}.$$

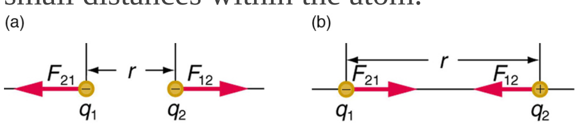
Coulomb's law calculates the magnitude of the force F between two point charges, q_1 and q_2, separated by a distance r. In SI units, the constant k is equal to

Equation:

$$k = 8.988 \times 10^9 \frac{\mathrm{N \cdot m^2}}{\mathrm{C^2}} \approx 8.99 \times 10^9 \frac{\mathrm{N \cdot m^2}}{\mathrm{C^2}}.$$

The electrostatic force is a vector quantity and is expressed in units of newtons. The force is understood to be along the line joining the two charges. (See [link].)

Although the formula for Coulomb's law is simple, it was no mean task to prove it. The experiments Coulomb did, with the primitive equipment then available, were difficult. Modern experiments have verified Coulomb's law to great precision. For example, it has been shown that the force is inversely proportional to distance between two objects squared $\left(F \propto 1/r^2\right)$ to an accuracy of 1 part in 10^{16}. No exceptions have ever been found, even at the small distances within the atom.



The magnitude of the electrostatic force F between point charges q_1 and q_2 separated by a distance r is given by Coulomb's law. Note that Newton's third law (every force exerted creates an equal and opposite force) applies as usual—the force on q_1 is equal in magnitude and opposite in direction to the force it exerts on q_2. (a) Like charges. (b) Unlike charges.

Example:

How Strong is the Coulomb Force Relative to the Gravitational Force?

Compare the electrostatic force between an electron and proton separated by 0.530×10^{-10} m with the gravitational force between them. This distance is their average separation in a hydrogen atom.

Strategy

To compare the two forces, we first compute the electrostatic force using Coulomb's law, $F = k\frac{|q_1 q_2|}{r^2}$. We then calculate the gravitational force using Newton's universal law of

gravitation. Finally, we take a ratio to see how the forces compare in magnitude.
Solution
Entering the given and known information about the charges and separation of the electron and proton into the expression of Coulomb's law yields
Equation:

$$F = k\frac{|q_1 q_2|}{r^2}$$

Equation:

$$= \left(8.99 \times 10^9\ \mathrm{N \cdot m^2/C^2}\right) \times \frac{(1.60\times10^{-19}\ \mathrm{C})(1.60\times10^{-19}\ \mathrm{C})}{(0.530\times10^{-10}\ \mathrm{m})^2}$$

Thus the Coulomb force is
Equation:

$$F = 8.19 \times 10^{-8}\ \mathrm{N}.$$

The charges are opposite in sign, so this is an attractive force. This is a very large force for an electron—it would cause an acceleration of $8.99 \times 10^{22}\ \mathrm{m/s^2}$(verification is left as an end-of-section problem).The gravitational force is given by Newton's law of gravitation as:
Equation:

$$F_G = G\frac{mM}{r^2},$$

where $G = 6.67 \times 10^{-11}\ \mathrm{N \cdot m^2/kg^2}$. Here m and M represent the electron and proton masses, which can be found in the appendices. Entering values for the knowns yields
Equation:

$$F_G = (6.67 \times 10^{-11}\ \mathrm{N \cdot m^2/kg^2}) \times \frac{(9.11 \times 10^{-31}\ \mathrm{kg})(1.67 \times 10^{-27}\ \mathrm{kg})}{(0.530 \times 10^{-10}\ \mathrm{m})^2} = 3.61 \times 10^{-47}\ \mathrm{N}$$

This is also an attractive force, although it is traditionally shown as positive since gravitational force is always attractive. The ratio of the magnitude of the electrostatic force to gravitational force in this case is, thus,
Equation:

$$\frac{F}{F_G} = 2.27 \times 10^{39}.$$

Discussion
This is a remarkably large ratio! Note that this will be the ratio of electrostatic force to gravitational force for an electron and a proton at any distance (taking the ratio before entering numerical values shows that the distance cancels). This ratio gives some indication

of just how much larger the Coulomb force is than the gravitational force between two of the most common particles in nature.

As the example implies, gravitational force is completely negligible on a small scale, where the interactions of individual charged particles are important. On a large scale, such as between the Earth and a person, the reverse is true. Most objects are nearly electrically neutral, and so attractive and repulsive **Coulomb forces** nearly cancel. Gravitational force on a large scale dominates interactions between large objects because it is always attractive, while Coulomb forces tend to cancel.

Section Summary

- Frenchman Charles Coulomb was the first to publish the mathematical equation that describes the electrostatic force between two objects.
- Coulomb's law gives the magnitude of the force between point charges. It is
  Equation:

  $$F = k\frac{|q_1 q_2|}{r^2},$$

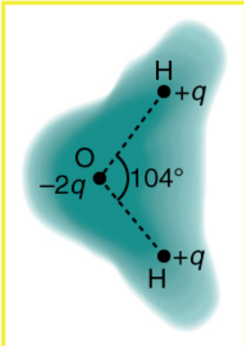
  where q_1 and q_2 are two point charges separated by a distance r, and $k \approx 8.99 \times 10^9\ \mathrm{N \cdot m^2/C^2}$
- This Coulomb force is extremely basic, since most charges are due to point-like particles. It is responsible for all electrostatic effects and underlies most macroscopic forces.
- The Coulomb force is extraordinarily strong compared with the gravitational force, another basic force—but unlike gravitational force it can cancel, since it can be either attractive or repulsive.
- The electrostatic force between two subatomic particles is far greater than the gravitational force between the same two particles.

Conceptual Questions

Exercise:

Problem:

[link] shows the charge distribution in a water molecule, which is called a polar molecule because it has an inherent separation of charge. Given water's polar character, explain what effect humidity has on removing excess charge from objects.



Schematic representation of the outer electron cloud of a neutral water molecule. The electrons spend more time near the oxygen than the hydrogens, giving a permanent charge separation as shown. Water is thus a *polar molecule*. It is more easily affected by electrostatic forces than molecules with uniform charge distributions.

Exercise:

Problem:

Using [link], explain, in terms of Coulomb's law, why a polar molecule (such as in [link]) is attracted by both positive and negative charges.

Exercise:

Problem:

Given the polar character of water molecules, explain how ions in the air form nucleation centers for rain droplets.

Problems & Exercises

Exercise:

Problem:

What is the repulsive force between two pith balls that are 8.00 cm apart and have equal charges of – 30.0 nC?

Exercise:

Problem:

(a) How strong is the attractive force between a glass rod with a $0.700\ \mu\text{C}$ charge and a silk cloth with a $-0.600\ \mu\text{C}$ charge, which are 12.0 cm apart, using the approximation that they act like point charges? (b) Discuss how the answer to this problem might be affected if the charges are distributed over some area and do not act like point charges.

Solution:

(a) 0.263 N

(b) If the charges are distributed over some area, there will be a concentration of charge along the side closest to the oppositely charged object. This effect will increase the net force.

Exercise:

Problem:

Two point charges exert a 5.00 N force on each other. What will the force become if the distance between them is increased by a factor of three?

Exercise:

Problem:

Two point charges are brought closer together, increasing the force between them by a factor of 25. By what factor was their separation decreased?

Solution:

The separation decreased by a factor of 5.

Exercise:

Problem:

How far apart must two point charges of 75.0 nC (typical of static electricity) be to have a force of 1.00 N between them?

Exercise:

Problem:

If two equal charges each of 1 C each are separated in air by a distance of 1 km, what is the magnitude of the force acting between them? You will see that even at a distance as large as 1 km, the repulsive force is substantial because 1 C is a very significant amount of charge.

Exercise:

Problem:

A test charge of $+2\ \mu\text{C}$ is placed halfway between a charge of $+6\ \mu\text{C}$ and another of $+4\ \mu\text{C}$ separated by 10 cm. (a) What is the magnitude of the force on the test charge? (b) What is the direction of this force (away from or toward the $+6\ \mu\text{C}$ charge)?

Exercise:

Problem:

Bare free charges do not remain stationary when close together. To illustrate this, calculate the acceleration of two isolated protons separated by 2.00 nm (a typical distance between gas atoms). Explicitly show how you follow the steps in the Problem-Solving Strategy for electrostatics.

Solution:

$$\begin{aligned} F &= k\frac{|q_1q_2|}{r^2} = ma \Rightarrow a = \frac{kq^2}{mr^2} \\ &= \frac{\left(9.00\times10^9\ \text{N}\cdot\text{m}^2/\text{C}^2\right)\left(1.60\times10^{-19}\ \text{m}\right)^2}{\left(1.67\times10^{-27}\ \text{kg}\right)\left(2.00\times10^{-9}\ \text{m}\right)^2} \\ &= 3.45\times10^{16}\ \text{m/s}^2 \end{aligned}$$

Exercise:

Problem:

(a) By what factor must you change the distance between two point charges to change the force between them by a factor of 10? (b) Explain how the distance can either increase or decrease by this factor and still cause a factor of 10 change in the force.

Solution:

(a) 3.2

(b) If the distance increases by 3.2, then the force will decrease by a factor of 10 ; if the distance decreases by 3.2, then the force will increase by a factor of 10. Either way, the force changes by a factor of 10.

Exercise:

Problem:

Suppose you have a total charge q_{tot} that you can split in any manner. Once split, the separation distance is fixed. How do you split the charge to achieve the greatest force?

Exercise:

Problem:

(a) Common transparent tape becomes charged when pulled from a dispenser. If one piece is placed above another, the repulsive force can be great enough to support the top piece's weight. Assuming equal point charges (only an approximation), calculate the magnitude of the charge if electrostatic force is great enough to support the weight of a 10.0 mg piece of tape held 1.00 cm above another. (b) Discuss whether the magnitude of this charge is consistent with what is typical of static electricity.

Solution:

(a) 1.04×10^{-9} C

(b) This charge is approximately 1 nC, which is consistent with the magnitude of charge typical for static electricity

Exercise:

Problem:

(a) Find the ratio of the electrostatic to gravitational force between two electrons. (b) What is this ratio for two protons? (c) Why is the ratio different for electrons and protons?

Exercise:

Problem:

At what distance is the electrostatic force between two protons equal to the weight of one proton?

Exercise:

Problem:

A certain five cent coin contains 5.00 g of nickel. What fraction of the nickel atoms' electrons, removed and placed 1.00 m above it, would support the weight of this coin? The atomic mass of nickel is 58.7, and each nickel atom contains 28 electrons and 28 protons.

Solution:

1.02×10^{-11}

Exercise:

Problem:

(a) Two point charges totaling $8.00\ \mu\mathrm{C}$ exert a repulsive force of 0.150 N on one another when separated by 0.500 m. What is the charge on each? (b) What is the charge on each if the force is attractive?

Exercise:

Problem:

Point charges of $5.00\ \mu\mathrm{C}$ and –$3.00\ \mu\mathrm{C}$ are placed 0.250 m apart. (a) Where can a third charge be placed so that the net force on it is zero? (b) What if both charges are positive?

Solution:

a. 0.859 m beyond negative charge on line connecting two charges
b. 0.109 m from lesser charge on line connecting two charges

Exercise:

Problem:

Two point charges q_1 and q_2 are $3.00\ \mathrm{m}$ apart, and their total charge is $20\ \mu\mathrm{C}$. (a) If the force of repulsion between them is 0.075N, what are magnitudes of the two charges? (b) If one charge attracts the other with a force of 0.525N, what are the magnitudes of the two charges? Note that you may need to solve a quadratic equation to reach your answer.

Glossary

Coulomb's law
: the mathematical equation calculating the electrostatic force vector between two charged particles

Coulomb force
: another term for the electrostatic force

electrostatic force

the amount and direction of attraction or repulsion between two charged bodies

Electric Field: Concept of a Field Revisited

- Describe a force field and calculate the strength of an electric field due to a point charge.
- Calculate the force exerted on a test charge by an electric field.
- Explain the relationship between electrical force (F) on a test charge and electrical field strength (E).

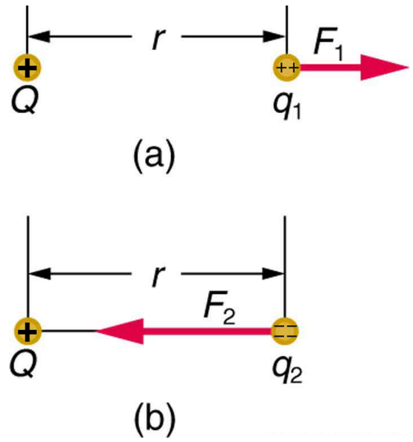
Contact forces, such as between a baseball and a bat, are explained on the small scale by the interaction of the charges in atoms and molecules in close proximity. They interact through forces that include the **Coulomb force**. Action at a distance is a force between objects that are not close enough for their atoms to “touch.” That is, they are separated by more than a few atomic diameters.

For example, a charged rubber comb attracts neutral bits of paper from a distance via the Coulomb force. It is very useful to think of an object being surrounded in space by a **force field**. The force field carries the force to another object (called a test object) some distance away.

Concept of a Field

A field is a way of conceptualizing and mapping the force that surrounds any object and acts on another object at a distance without apparent physical connection. For example, the gravitational field surrounding the earth (and all other masses) represents the gravitational force that would be experienced if another mass were placed at a given point within the field.

In the same way, the Coulomb force field surrounding any charge extends throughout space. Using Coulomb’s law, $F=k|q_1q_2|/r^2$, its magnitude is given by the equation $F=k|\,qQ\,|/r^2$, for a **point charge** (a particle having a charge Q) acting on a **test charge** q at a distance r (see [link]). Both the magnitude and direction of the Coulomb force field depend on Q and the test charge q.



The Coulomb force field due to a positive charge Q is shown acting on two different charges. Both charges are the same distance from Q. (a) Since q_1 is positive, the force F_1 acting on it is repulsive. (b) The charge q_2 is negative and greater in magnitude than q_1, and so the force F_2 acting on it is attractive and stronger than F_1. The Coulomb force field is thus not unique at any point in space, because it depends on the test charges q_1 and q_2

as well as the
charge Q.

To simplify things, we would prefer to have a field that depends only on Q and not on the test charge q. The electric field is defined in such a manner that it represents only the charge creating it and is unique at every point in space. Specifically, the electric field E is defined to be the ratio of the Coulomb force to the test charge:
Equation:

$$\mathbf{E} = \frac{\mathbf{F}}{q},$$

where $\mathbf{F}$ is the electrostatic force (or Coulomb force) exerted on a positive test charge q. It is understood that $\mathbf{E}$ is in the same direction as $\mathbf{F}$. It is also assumed that q is so small that it does not alter the charge distribution creating the electric field. The units of electric field are newtons per coulomb (N/C). If the electric field is known, then the electrostatic force on any charge q is simply obtained by multiplying charge times electric field, or $\mathbf{F} = q\mathbf{E}$. Consider the electric field due to a point charge Q. According to Coulomb's law, the force it exerts on a test charge q is $F = k|\, qQ \,|/r^2$. Thus the magnitude of the electric field, E, for a point charge is
Equation:

$$E = \left|\frac{F}{q}\right| = k\left|\frac{\text{qQ}}{qr^2}\right| = k\frac{|Q|}{r^2}.$$

Since the test charge cancels, we see that
Equation:

$$E = k\frac{|Q|}{r^2}.$$

The electric field is thus seen to depend only on the charge Q and the distance r; it is completely independent of the test charge q.

Example:
Calculating the Electric Field of a Point Charge
Calculate the strength and direction of the electric field E due to a point charge of 2.00 nC (nano-Coulombs) at a distance of 5.00 mm from the charge.
Strategy
We can find the electric field created by a point charge by using the equation $E = \mathrm{kQ}/r^2$.
Solution
Here $Q = 2.00 \times 10^{-9}$ C and $r = 5.00 \times 10^{-3}$ m. Entering those values into the above equation gives
Equation:

$$\begin{aligned} E &= k\frac{Q}{r^2} \\ &= (8.99 \times 10^9\ \mathrm{N \cdot m^2/C^2}) \times \frac{(2.00\times10^{-9}\ \mathrm{C})}{(5.00\times10^{-3}\ \mathrm{m})^2} \\ &= 7.19 \times 10^5\ \mathrm{N/C}. \end{aligned}$$

Discussion
This **electric field strength** is the same at any point 5.00 mm away from the charge Q that creates the field. It is positive, meaning that it has a direction pointing away from the charge Q.

Example:
Calculating the Force Exerted on a Point Charge by an Electric Field
What force does the electric field found in the previous example exert on a point charge of –0.250 μC?
Strategy
Since we know the electric field strength and the charge in the field, the force on that charge can be calculated using the definition of electric field

$\mathbf{E} = \mathbf{F}/q$ rearranged to $\mathbf{F} = q\mathbf{E}$.
Solution
The magnitude of the force on a charge $q = -0.250\ \mu\text{C}$ exerted by a field of strength $E = 7.20 \times 10^5$ N/C is thus,
Equation:

$$\begin{array}{rcl} F & = & -qE \\ & = & (0.250 \times 10^{-6}\ \text{C})(7.20 \times 10^5\ \text{N/C}) \\ & = & 0.180\ \text{N}. \end{array}$$

Because q is negative, the force is directed opposite to the direction of the field.
Discussion
The force is attractive, as expected for unlike charges. (The field was created by a positive charge and here acts on a negative charge.) The charges in this example are typical of common static electricity, and the modest attractive force obtained is similar to forces experienced in static cling and similar situations.

Note:
PhET Explorations: Electric Field of Dreams
Play ball! Add charges to the Field of Dreams and see how they react to the electric field. Turn on a background electric field and adjust the direction and magnitude.
https://archive.cnx.org/specials/ca9a78b4-06a7-11e6-b638-3bb71d1f0b42/electric-field-of-dreams/#sim-electric-field-of-dreams

Section Summary

- The electrostatic force field surrounding a charged object extends out into space in all directions.
- The electrostatic force exerted by a point charge on a test charge at a distance r depends on the charge of both charges, as well as the

distance between the two.

- The electric field $\mathbf{E}$ is defined to be
  Equation:
  $$\mathbf{E} = \frac{\mathbf{F}}{q,}$$
  where $\mathbf{F}$ is the Coulomb or electrostatic force exerted on a small positive test charge q. $\mathbf{E}$ has units of N/C.
- The magnitude of the electric field $\mathbf{E}$ created by a point charge Q is
  Equation:
  $$\mathbf{E} = k\frac{|Q|}{r^2}.$$
  where r is the distance from Q. The electric field $\mathbf{E}$ is a vector and fields due to multiple charges add like vectors.

Conceptual Questions

Exercise:

Problem:

Why must the test charge q in the definition of the electric field be vanishingly small?

Exercise:

Problem:

Are the direction and magnitude of the Coulomb force unique at a given point in space? What about the electric field?

Problem Exercises

Exercise:

Problem:

What is the magnitude and direction of an electric field that exerts a 2.00×10^{-5} N upward force on a –$1.75\ \mu$C charge?

Exercise:

Problem:

What is the magnitude and direction of the force exerted on a $3.50\ \mu$C charge by a 250 N/C electric field that points due east?

Solution:

8.75×10^{-4} N

Exercise:

Problem:

Calculate the magnitude of the electric field 2.00 m from a point charge of 5.00 mC (such as found on the terminal of a Van de Graaff).

Exercise:

Problem:

(a) What magnitude point charge creates a 10,000 N/C electric field at a distance of 0.250 m? (b) How large is the field at 10.0 m?

Solution:

(a) 6.94×10^{-8} C

(b) $6.25\ \mathrm{N/C}$

Exercise:

Problem:

Calculate the initial (from rest) acceleration of a proton in a $5.00 \times 10^6\ \mathrm{N/C}$ electric field (such as created by a research Van de Graaff). Explicitly show how you follow the steps in the Problem-Solving Strategy for electrostatics.

Exercise:

Problem:

(a) Find the magnitude and direction of an electric field that exerts a $4.80 \times 10^{-17}\ \mathrm{N}$ westward force on an electron. (b) What magnitude and direction force does this field exert on a proton?

Solution:

(a) $300\ \mathrm{N/C}\ (\mathrm{east})$

(b) $4.80 \times 10^{-17}\ \mathrm{N}\ (\mathrm{east})$

Glossary

field
: a map of the amount and direction of a force acting on other objects, extending out into space

point charge
: A charged particle, designated Q, generating an electric field

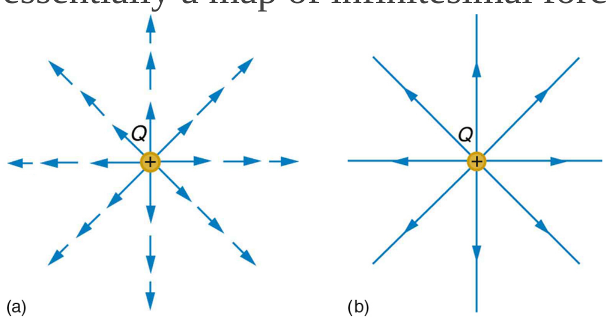
test charge
: A particle (designated q) with either a positive or negative charge set down within an electric field generated by a point charge

Electric Field Lines: Multiple Charges

- Calculate the total force (magnitude and direction) exerted on a test charge from more than one charge
- Describe an electric field diagram of a positive point charge; of a negative point charge with twice the magnitude of positive charge
- Draw the electric field lines between two points of the same charge; between two points of opposite charge.

Drawings using lines to represent **electric fields** around charged objects are very useful in visualizing field strength and direction. Since the electric field has both magnitude and direction, it is a vector. Like all **vectors**, the electric field can be represented by an arrow that has length proportional to its magnitude and that points in the correct direction. (We have used arrows extensively to represent force vectors, for example.)

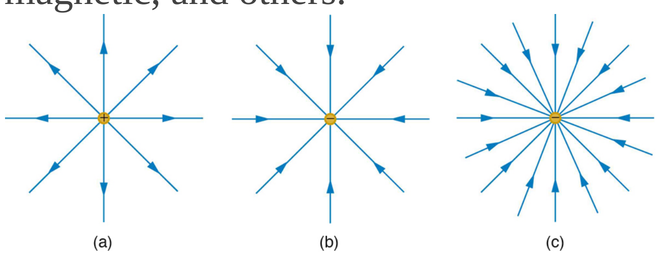
[link] shows two pictorial representations of the same electric field created by a positive point charge Q. [link] (b) shows the standard representation using continuous lines. [link] (a) shows numerous individual arrows with each arrow representing the force on a test charge q. Field lines are essentially a map of infinitesimal force vectors.



Two equivalent representations of the electric field due to a positive charge Q. (a) Arrows representing the electric field's magnitude and direction. (b) In the standard representation, the arrows are replaced by continuous field lines having the same direction at any point

as the electric field. The closeness of the lines is directly related to the strength of the electric field. A test charge placed anywhere will feel a force in the direction of the field line; this force will have a strength proportional to the density of the lines (being greater near the charge, for example).

Note that the electric field is defined for a positive test charge q, so that the field lines point away from a positive charge and toward a negative charge. (See [link].) The electric field strength is exactly proportional to the number of field lines per unit area, since the magnitude of the electric field for a point charge is $E = k|Q|/r^2$ and area is proportional to r^2. This pictorial representation, in which field lines represent the direction and their closeness (that is, their areal density or the number of lines crossing a unit area) represents strength, is used for all fields: electrostatic, gravitational, magnetic, and others.

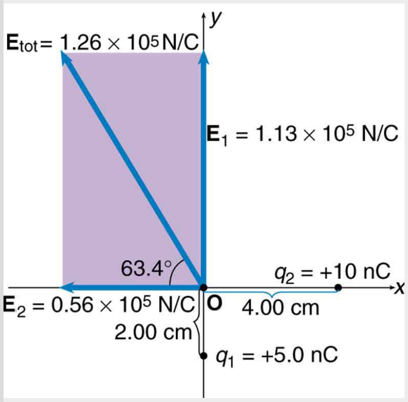


The electric field surrounding three different point charges. (a) A positive charge. (b) A negative charge of equal magnitude. (c) A larger negative charge.

In many situations, there are multiple charges. The total electric field created by multiple charges is the vector sum of the individual fields created by each charge. The following example shows how to add electric field vectors.

Example:
Adding Electric Fields

Find the magnitude and direction of the total electric field due to the two point charges, q_1 and q_2, at the origin of the coordinate system as shown in [link].



The electric fields $\mathbf{E}_1$ and $\mathbf{E}_2$ at the origin O add to $\mathbf{E}_{\text{tot}}$.

Strategy

Since the electric field is a vector (having magnitude and direction), we add electric fields with the same vector techniques used for other types of vectors. We first must find the electric field due to each charge at the point of interest, which is the origin of the coordinate system (O) in this instance. We pretend that there is a positive test charge, q, at point O, which allows us to determine the direction of the fields $\mathbf{E}_1$ and $\mathbf{E}_2$. Once those fields are found, the total field can be determined using **vector addition**.

Solution

The electric field strength at the origin due to q_1 is labeled E_1 and is calculated:

Equation:

$$E_1 = k\frac{q_1}{r_1^2} = \left(8.99 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2\right)\frac{\left(5.00\times10^{-9}\text{ C}\right)}{\left(2.00\times10^{-2}\text{ m}\right)^2}$$

$$E_1 = 1.124 \times 10^5 \text{ N/C}.$$

Similarly, E_2 is

Equation:

$$E_2 = k\frac{q_2}{r_2^2} = \left(8.99 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2\right)\frac{\left(10.0\times10^{-9}\text{ C}\right)}{\left(4.00\times10^{-2}\text{ m}\right)^2}$$

$$E_2 = 0.5619 \times 10^5 \text{ N/C}.$$

Four digits have been retained in this solution to illustrate that E_1 is exactly twice the magnitude of E_2. Now arrows are drawn to represent the magnitudes and directions of $\mathbf{E}_1$ and $\mathbf{E}_2$. (See [link].) The direction of the electric field is that of the force on a positive charge so both arrows point directly away from the positive charges that create them. The arrow for $\mathbf{E}_1$ is exactly twice the length of that for $\mathbf{E}_2$. The arrows form a right triangle in this case and can be added using the Pythagorean theorem. The magnitude of the total field E_{tot} is

Equation:

$$\begin{aligned} E_{\text{tot}} &= (E_1^2 + E_2^2)^{1/2} \\ &= \left\{(1.124 \times 10^5 \text{ N/C})^2 + (0.5619 \times 10^5 \text{ N/C})^2\right\}^{1/2} \\ &= 1.26 \times 10^5 \text{ N/C}. \end{aligned}$$

The direction is

Equation:

$$\begin{aligned}\theta &= \tan^{-1}\left(\frac{E_1}{E_2}\right)\\ &= \tan^{-1}\left(\frac{1.124\times10^5\text{ N/C}}{0.5619\times10^5\text{ N/C}}\right)\\ &= 63.4^{\text{o}},\end{aligned}$$

or 63.4^{o} above the *x*-axis.

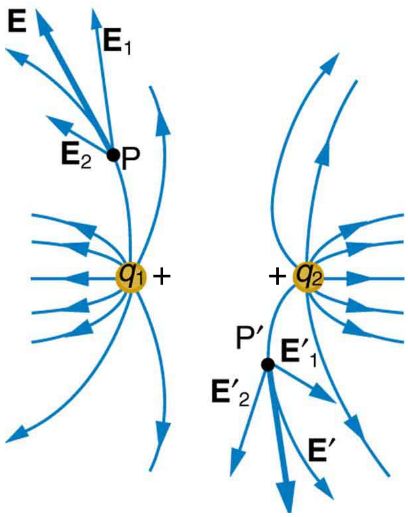
Discussion

In cases where the electric field vectors to be added are not perpendicular, vector components or graphical techniques can be used. The total electric field found in this example is the total electric field at only one point in space. To find the total electric field due to these two charges over an entire region, the same technique must be repeated for each point in the region. This impossibly lengthy task (there are an infinite number of points in space) can be avoided by calculating the total field at representative points and using some of the unifying features noted next.

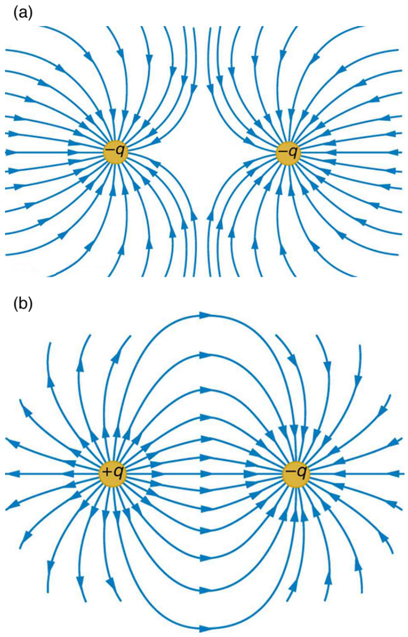
[link] shows how the electric field from two point charges can be drawn by finding the total field at representative points and drawing electric field lines consistent with those points. While the electric fields from multiple charges are more complex than those of single charges, some simple features are easily noticed.

For example, the field is weaker between like charges, as shown by the lines being farther apart in that region. (This is because the fields from each charge exert opposing forces on any charge placed between them.) (See [link] and [link](a).) Furthermore, at a great distance from two like charges, the field becomes identical to the field from a single, larger charge.

[link](b) shows the electric field of two unlike charges. The field is stronger between the charges. In that region, the fields from each charge are in the same direction, and so their strengths add. The field of two unlike charges is weak at large distances, because the fields of the individual charges are in opposite directions and so their strengths subtract. At very large distances, the field of two unlike charges looks like that of a smaller single charge.



Two positive point charges q_1 and q_2 produce the resultant electric field shown. The field is calculated at representative points and then smooth field lines drawn following the rules outlined in the text.



(a) Two negative charges produce the fields shown. It is very similar to the field produced by two positive charges, except that the directions are reversed. The field is clearly weaker between the charges. The individual forces on a test charge in that region are in opposite directions. (b) Two opposite charges produce the field shown, which is stronger in the region between the charges.

We use electric field lines to visualize and analyze electric fields (the lines are a pictorial tool, not a physical entity in themselves). The properties of electric field lines for any charge distribution can be summarized as follows:

1. Field lines must begin on positive charges and terminate on negative charges, or at infinity in the hypothetical case of isolated charges.
2. The number of field lines leaving a positive charge or entering a negative charge is proportional to the magnitude of the charge.
3. The strength of the field is proportional to the closeness of the field lines—more precisely, it is proportional to the number of lines per unit area perpendicular to the lines.
4. The direction of the electric field is tangent to the field line at any point in space.
5. Field lines can never cross.

The last property means that the field is unique at any point. The field line represents the direction of the field; so if they crossed, the field would have two directions at that location (an impossibility if the field is unique).

Note:
PhET Explorations: Charges and Fields
Move point charges around on the playing field and then view the electric field, voltages, equipotential lines, and more. It's colorful, it's dynamic, it's free.

Click here for the simulation.

Section Summary

- Drawings of electric field lines are useful visual tools. The properties of electric field lines for any charge distribution are that:
- Field lines must begin on positive charges and terminate on negative charges, or at infinity in the hypothetical case of isolated charges.
- The number of field lines leaving a positive charge or entering a negative charge is proportional to the magnitude of the charge.
- The strength of the field is proportional to the closeness of the field lines—more precisely, it is proportional to the number of lines per unit area perpendicular to the lines.
- The direction of the electric field is tangent to the field line at any point in space.
- Field lines can never cross.

Conceptual Questions

Exercise:

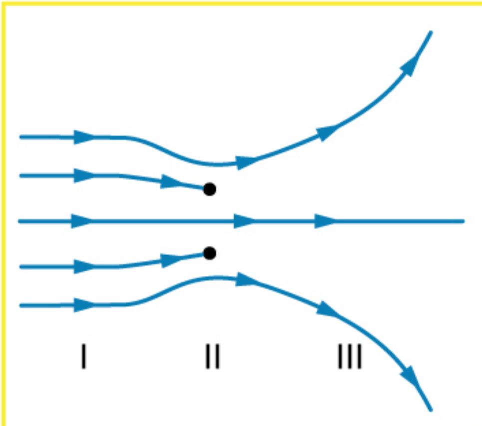
Problem:

Compare and contrast the Coulomb force field and the electric field. To do this, make a list of five properties for the Coulomb force field analogous to the five properties listed for electric field lines. Compare each item in your list of Coulomb force field properties with those of the electric field—are they the same or different? (For example, electric field lines cannot cross. Is the same true for Coulomb field lines?)

Exercise:

Problem:

[link] shows an electric field extending over three regions, labeled I, II, and III. Answer the following questions. (a) Are there any isolated charges? If so, in what region and what are their signs? (b) Where is the field strongest? (c) Where is it weakest? (d) Where is the field the most uniform?



Problem Exercises

Exercise:

Problem:

(a) Sketch the electric field lines near a point charge $+q$. (b) Do the same for a point charge $-3.00q$.

Exercise:

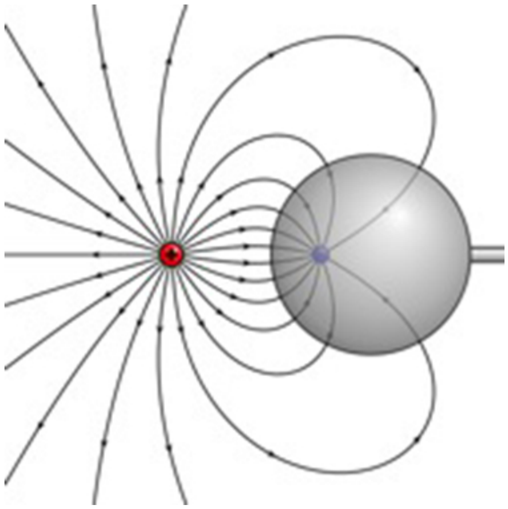
Problem:

Sketch the electric field lines a long distance from the charge distributions shown in [link] (a) and (b)

Exercise:

Problem:

[link] shows the electric field lines near two charges q_1 and q_2. What is the ratio of their magnitudes? (b) Sketch the electric field lines a long distance from the charges shown in the figure.



The electric field near two charges.

Exercise:

Problem:

Sketch the electric field lines in the vicinity of two opposite charges, where the negative charge is three times greater in magnitude than the positive. (See [link] for a similar situation).

Glossary

electric field
: a three-dimensional map of the electric force extended out into space from a point charge

electric field lines
: a series of lines drawn from a point charge representing the magnitude and direction of force exerted by that charge

vector
: a quantity with both magnitude and direction

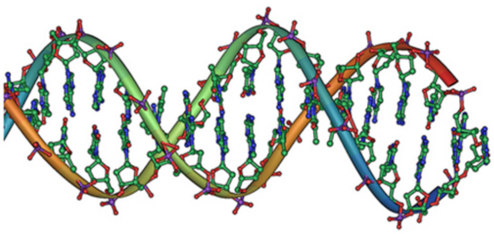
vector addition

mathematical combination of two or more vectors, including their magnitudes, directions, and positions

Electric Forces in Biology

- Describe how a water molecule is polar.
- Explain electrostatic screening by a water molecule within a living cell.

Classical electrostatics has an important role to play in modern molecular biology. Large molecules such as proteins, nucleic acids, and so on—so important to life—are usually electrically charged. DNA itself is highly charged; it is the electrostatic force that not only holds the molecule together but gives the molecule structure and strength. [link] is a schematic of the DNA double helix.



DNA is a highly charged molecule. The DNA double helix shows the two coiled strands each containing a row of nitrogenous bases, which “code” the genetic information needed by a living organism. The strands are connected by bonds between pairs of bases. While pairing combinations between certain bases are fixed (C-G and A-T), the sequence of nucleotides in the strand varies. (credit: Jerome Walker)

The four nucleotide bases are given the symbols A (adenine), C (cytosine), G (guanine), and T (thymine). The order of the four bases varies in each strand, but the pairing between bases is always the same. C and G are always paired and A and T are always paired, which helps to preserve the order of bases in cell division (mitosis) so as to pass on the correct genetic information. Since the Coulomb force drops with distance ($F \propto 1/r^2$), the distances between the base pairs must be small enough that the electrostatic force is sufficient to hold them together.

DNA is a highly charged molecule, with about $2q_\mathrm{e}$ (fundamental charge) per 0.3×10^{-9} m. The distance separating the two strands that make up the DNA structure is about 1 nm, while the distance separating the individual atoms within each base is about 0.3 nm.

One might wonder why electrostatic forces do not play a larger role in biology than they do if we have so many charged molecules. The reason is that the electrostatic force is "diluted" due to **screening** between molecules. This is due to the presence of other charges in the cell.

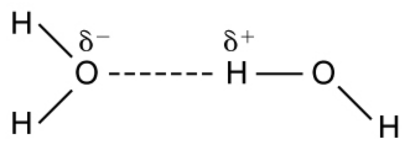
Polarity of Water Molecules

The best example of this charge screening is the water molecule, represented as $\mathrm{H_2O}$. Water is a strongly **polar molecule**. Its 10 electrons (8 from the oxygen atom and 2 from the two hydrogen atoms) tend to remain closer to the oxygen nucleus than the hydrogen nuclei. This creates two centers of equal and opposite charges—what is called a **dipole**, as illustrated in [link]. The magnitude of the dipole is called the dipole moment.

These two centers of charge will terminate some of the electric field lines coming from a free charge, as on a DNA molecule. This results in a reduction in the strength of the **Coulomb interaction**. One might say that screening makes the Coulomb force a short range force rather than long range.

Other ions of importance in biology that can reduce or screen Coulomb interactions are Na^+, and K^+, and Cl^-. These ions are located both inside and outside of living cells. The movement of these ions through cell membranes is crucial to the motion of nerve impulses through nerve axons.

Recent studies of electrostatics in biology seem to show that electric fields in cells can be extended over larger distances, in spite of screening, by "microtubules" within the cell. These microtubules are hollow tubes composed of proteins that guide the movement of chromosomes when cells divide, the motion of other organisms within the cell, and provide mechanisms for motion of some cells (as motors).



This schematic shows water (H_2O) as a polar molecule. Unequal sharing of electrons between the oxygen (O) and hydrogen (H) atoms leads to a net separation of positive and negative charge—forming a dipole. The symbols δ^- and δ^+ indicate that the oxygen side of the H_2O molecule tends to be more negative, while the hydrogen ends tend

to be more positive. This leads to an attraction of opposite charges between molecules.

Section Summary

- Many molecules in living organisms, such as DNA, carry a charge.
- An uneven distribution of the positive and negative charges within a polar molecule produces a dipole.
- The effect of a Coulomb field generated by a charged object may be reduced or blocked by other nearby charged objects.
- Biological systems contain water, and because water molecules are polar, they have a strong effect on other molecules in living systems.

Conceptual Question

Exercise:

Problem:

A cell membrane is a thin layer enveloping a cell. The thickness of the membrane is much less than the size of the cell. In a static situation the membrane has a charge distribution of -2.5×10^{-6}C/m ^{2} on its inner surface and $+2.5 \times 10^{-6}$ C/m^2 on its outer surface. Draw a diagram of the cell and the surrounding cell membrane. Include on this diagram the charge distribution and the corresponding electric field. Is there any electric field inside the cell? Is there any electric field outside the cell?

Glossary

dipole

a molecule’s lack of symmetrical charge distribution, causing one side to be more positive and another to be more negative

polar molecule
a molecule with an asymmetrical distribution of positive and negative charge

screening
the dilution or blocking of an electrostatic force on a charged object by the presence of other charges nearby

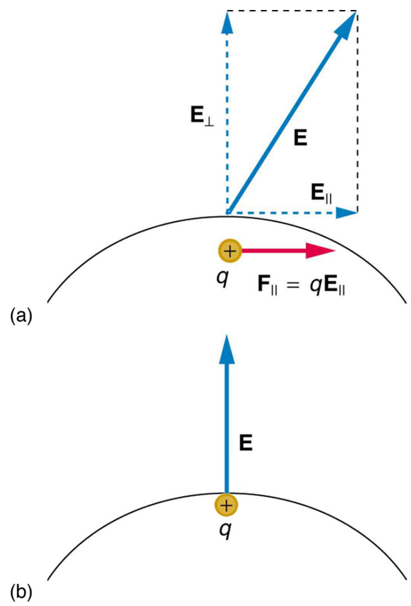
Coulomb interaction
the interaction between two charged particles generated by the Coulomb forces they exert on one another

Conductors and Electric Fields in Static Equilibrium

- List the three properties of a conductor in electrostatic equilibrium.
- Explain the effect of an electric field on free charges in a conductor.
- Explain why no electric field may exist inside a conductor.
- Describe the electric field surrounding Earth.
- Explain what happens to an electric field applied to an irregular conductor.
- Describe how a lightning rod works.
- Explain how a metal car may protect passengers inside from the dangerous electric fields caused by a downed line touching the car.

Conductors contain **free charges** that move easily. When excess charge is placed on a conductor or the conductor is put into a static electric field, charges in the conductor quickly respond to reach a steady state called **electrostatic equilibrium**.

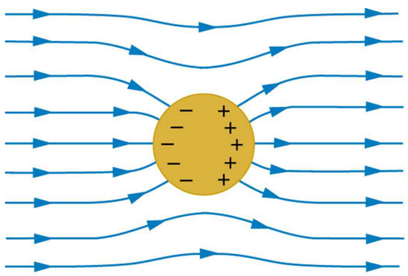
[link] shows the effect of an electric field on free charges in a conductor. The free charges move until the field is perpendicular to the conductor’s surface. There can be no component of the field parallel to the surface in electrostatic equilibrium, since, if there were, it would produce further movement of charge. A positive free charge is shown, but free charges can be either positive or negative and are, in fact, negative in metals. The motion of a positive charge is equivalent to the motion of a negative charge in the opposite direction.



When an electric field $\mathbf{E}$ is applied to a conductor, free charges inside the conductor move until the field is perpendicular to the surface. (a) The electric field is a vector quantity, with both parallel and perpendicular components. The parallel component $(\mathbf{E}_{||})$ exerts a force $(\mathbf{F}_{||})$ on the free charge q, which moves the charge until $\mathbf{F}_{||} = 0$. (b) The resulting field is perpendicular to the surface. The free charge has

been brought to the conductor’s surface, leaving electrostatic forces in equilibrium.

A conductor placed in an **electric field** will be **polarized**. [link] shows the result of placing a neutral conductor in an originally uniform electric field. The field becomes stronger near the conductor but entirely disappears inside it.



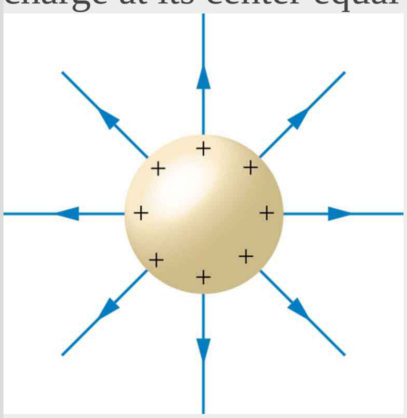
This illustration shows a spherical conductor in static equilibrium with an originally uniform electric field. Free charges move within the conductor, polarizing it, until the electric field lines are perpendicular to the surface. The field lines end on excess negative charge on one section of the surface and begin

again on excess positive charge on the opposite side. No electric field exists inside the conductor, since free charges in the conductor would continue moving in response to any field until it was neutralized.

Note:

Misconception Alert: Electric Field inside a Conductor

Excess charges placed on a spherical conductor repel and move until they are evenly distributed, as shown in [link]. Excess charge is forced to the surface until the field inside the conductor is zero. Outside the conductor, the field is exactly the same as if the conductor were replaced by a point charge at its center equal to the excess charge.



The mutual repulsion of excess positive charges on

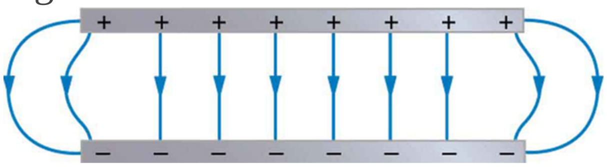
a spherical conductor distributes them uniformly on its surface. The resulting electric field is perpendicular to the surface and zero inside. Outside the conductor, the field is identical to that of a point charge at the center equal to the excess charge.

Note:
Properties of a Conductor in Electrostatic Equilibrium

1. The electric field is zero inside a conductor.
2. Just outside a conductor, the electric field lines are perpendicular to its surface, ending or beginning on charges on the surface.
3. Any excess charge resides entirely on the surface or surfaces of a conductor.

The properties of a conductor are consistent with the situations already discussed and can be used to analyze any conductor in electrostatic equilibrium. This can lead to some interesting new insights, such as described below.

How can a very uniform electric field be created? Consider a system of two metal plates with opposite charges on them, as shown in [link]. The properties of conductors in electrostatic equilibrium indicate that the electric field between the plates will be uniform in strength and direction. Except near the edges, the excess charges distribute themselves uniformly, producing field lines that are uniformly spaced (hence uniform in strength) and perpendicular to the surfaces (hence uniform in direction, since the plates are flat). The edge effects are less important when the plates are close together.



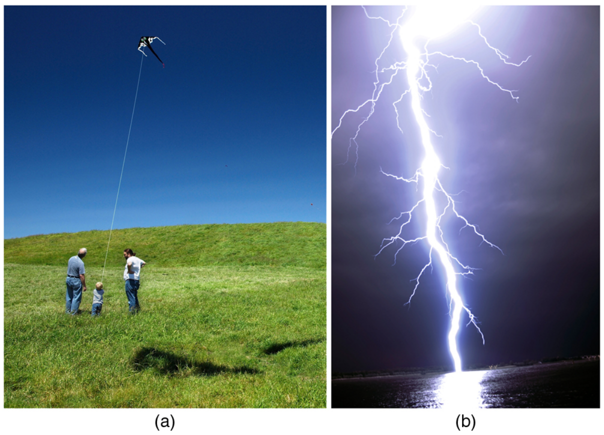
Two metal plates with equal, but opposite, excess charges. The field between them is uniform in strength and direction except near the edges. One use of such a field is to produce uniform acceleration of charges between the plates, such as in the electron gun of a TV tube.

Earth's Electric Field

A near uniform electric field of approximately 150 N/C, directed downward, surrounds Earth, with the magnitude increasing slightly as we get closer to the surface. What causes the electric field? At around 100 km above the surface of Earth we have a layer of charged particles, called the **ionosphere**. The ionosphere is responsible for a range of phenomena including the electric field surrounding Earth. In fair weather the ionosphere is positive and the Earth largely negative, maintaining the electric field ([link](a)).

In storm conditions clouds form and localized electric fields can be larger and reversed in direction ([link](b)). The exact charge distributions depend on the local conditions, and variations of [link](b) are possible.

If the electric field is sufficiently large, the insulating properties of the surrounding material break down and it becomes conducting. For air this occurs at around 3×10^6 N/C. Air ionizes ions and electrons recombine, and we get discharge in the form of lightning sparks and corona discharge.



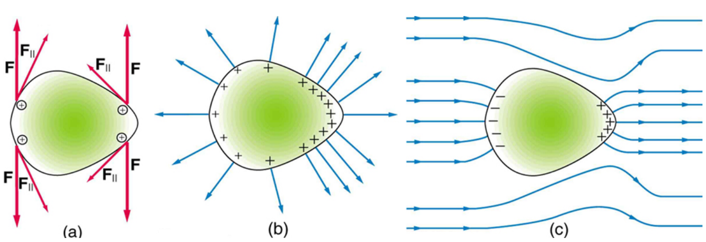
Earth's electric field. (a) Fair weather field. Earth and the ionosphere (a layer of charged particles) are both conductors. They produce a uniform electric field of about 150 N/C. (credit: D. H. Parks) (b) Storm fields. In the presence of storm clouds, the local electric fields can be larger. At very high fields, the insulating properties of the air break down and lightning can occur. (credit: Jan-Joost Verhoef)

Electric Fields on Uneven Surfaces

So far we have considered excess charges on a smooth, symmetrical conductor surface. What happens if a conductor has sharp corners or is pointed? Excess charges on a nonuniform conductor become concentrated at the sharpest points. Additionally, excess charge may move on or off the conductor at the sharpest points.

To see how and why this happens, consider the charged conductor in [link]. The electrostatic repulsion of like charges is most effective in moving them apart on the flattest surface, and so they become least concentrated there. This is because the forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surfaces are different. The component parallel to the surface is greatest on the flattest surface and, hence, more effective in moving the charge.

The same effect is produced on a conductor by an externally applied electric field, as seen in [link] (c). Since the field lines must be perpendicular to the surface, more of them are concentrated on the most curved parts.



Excess charge on a nonuniform conductor becomes most concentrated at the location of greatest curvature. (a) The forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surface are different. It is $\mathbf{F}_{||}$ that moves the charges apart once they

have reached the surface. (b) $\mathbf{F}_{||}$ is smallest at the more pointed end, the charges are left closer together, producing the electric field shown. (c) An uncharged conductor in an originally uniform electric field is polarized, with the most concentrated charge at its most pointed end.

Applications of Conductors

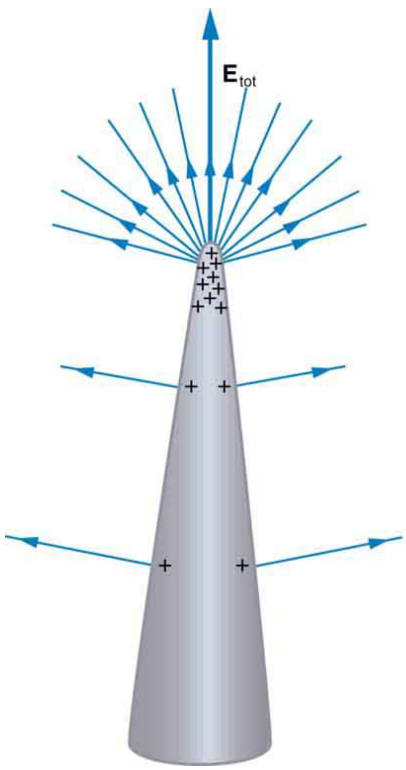
On a very sharply curved surface, such as shown in [link], the charges are so concentrated at the point that the resulting electric field can be great enough to remove them from the surface. This can be useful.

Lightning rods work best when they are most pointed. The large charges created in storm clouds induce an opposite charge on a building that can result in a lightning bolt hitting the building. The induced charge is bled away continually by a lightning rod, preventing the more dramatic lightning strike.

Of course, we sometimes wish to prevent the transfer of charge rather than to facilitate it. In that case, the conductor should be very smooth and have as large a radius of curvature as possible. (See [link].) Smooth surfaces are used on high-voltage transmission lines, for example, to avoid leakage of charge into the air.

Another device that makes use of some of these principles is a **Faraday cage**. This is a metal shield that encloses a volume. All electrical charges will reside on the outside surface of this shield, and there will be no electrical field inside. A Faraday cage is used to prohibit stray electrical fields in the environment from interfering with sensitive measurements, such as the electrical signals inside a nerve cell.

During electrical storms if you are driving a car, it is best to stay inside the car as its metal body acts as a Faraday cage with zero electrical field inside. If in the vicinity of a lightning strike, its effect is felt on the outside of the car and the inside is unaffected, provided you remain totally inside. This is also true if an active (“hot”) electrical wire was broken (in a storm or an accident) and fell on your car.



A very pointed conductor has a large charge concentration at the point. The electric field is very strong at the point and can exert a force large enough to transfer charge on or off the conductor.

Lightning rods are used to prevent the buildup of large excess charges on structures and, thus, are pointed.

(a) A lightning rod is pointed to facilitate the transfer of charge. (credit: Romaine, Wikimedia Commons) (b) This Van de Graaff generator has a smooth surface with a large radius of curvature to prevent the transfer of charge and allow a large voltage to be generated. The mutual repulsion of like charges is evident in the person’s hair while touching the metal sphere. (credit: Jon ‘ShakataGaNai’ Davis/Wikimedia Commons).

Section Summary

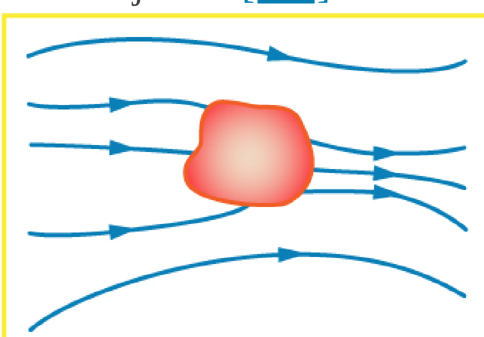
- A conductor allows free charges to move about within it.
- The electrical forces around a conductor will cause free charges to move around inside the conductor until static equilibrium is reached.
- Any excess charge will collect along the surface of a conductor.
- Conductors with sharp corners or points will collect more charge at those points.
- A lightning rod is a conductor with sharply pointed ends that collect excess charge on the building caused by an electrical storm and allow it to dissipate back into the air.
- Electrical storms result when the electrical field of Earth's surface in certain locations becomes more strongly charged, due to changes in the insulating effect of the air.
- A Faraday cage acts like a shield around an object, preventing electric charge from penetrating inside.

Conceptual Questions

Exercise:

Problem:

Is the object in [link] a conductor or an insulator? Justify your answer.



Exercise:

Problem:

If the electric field lines in the figure above were perpendicular to the object, would it necessarily be a conductor? Explain.

Exercise:

Problem:

The discussion of the electric field between two parallel conducting plates, in this module states that edge effects are less important if the plates are close together. What does close mean? That is, is the actual plate separation crucial, or is the ratio of plate separation to plate area crucial?

Exercise:

Problem:

Would the self-created electric field at the end of a pointed conductor, such as a lightning rod, remove positive or negative charge from the conductor? Would the same sign charge be removed from a neutral pointed conductor by the application of a similar externally created electric field? (The answers to both questions have implications for charge transfer utilizing points.)

Exercise:

Problem:

Why is a golfer with a metal club over her shoulder vulnerable to lightning in an open fairway? Would she be any safer under a tree?

Exercise:

Problem:

Can the belt of a Van de Graaff accelerator be a conductor? Explain.

Exercise:

Problem:

Are you relatively safe from lightning inside an automobile? Give two reasons.

Exercise:

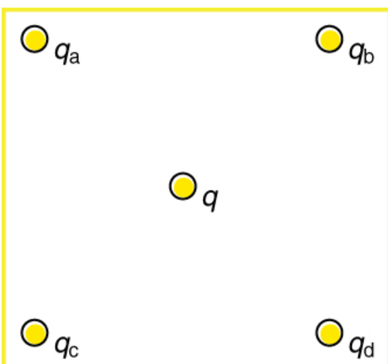
Problem:

Discuss pros and cons of a lightning rod being grounded versus simply being attached to a building.

Exercise:

Problem:

Using the symmetry of the arrangement, show that the net Coulomb force on the charge q at the center of the square below ([link]) is zero if the charges on the four corners are exactly equal.



Four point charges q_a, q_b, q_c, and q_d lie on the corners of a square and q is located at its center.

Exercise:

Problem:

(a) Using the symmetry of the arrangement, show that the electric field at the center of the square in [link] is zero if the charges on the four corners are exactly equal. (b) Show that this is also true for any combination of charges in which $q_a = q_d$ and $q_b = q_c$

Exercise:

Problem:

(a) What is the direction of the total Coulomb force on q in [link] if q is negative, $q_a = q_c$ and both are negative, and $q_b = q_c$ and both are positive? (b) What is the direction of the electric field at the center of the square in this situation?

Exercise:

Problem:

Considering [link], suppose that $q_a = q_d$ and $q_b = q_c$. First show that q is in static equilibrium. (You may neglect the gravitational force.) Then discuss whether the equilibrium is stable or unstable, noting that this may depend on the signs of the charges and the direction of displacement of q from the center of the square.

Exercise:

Problem:

If $q_a = 0$ in [link], under what conditions will there be no net Coulomb force on q?

Exercise:

Problem:

In regions of low humidity, one develops a special “grip” when opening car doors, or touching metal door knobs. This involves placing as much of the hand on the device as possible, not just the ends of one’s fingers. Discuss the induced charge and explain why this is done.

Exercise:

Problem:

Tollbooth stations on roadways and bridges usually have a piece of wire stuck in the pavement before them that will touch a car as it approaches. Why is this done?

Exercise:

Problem:

Suppose a woman carries an excess charge. To maintain her charged status can she be standing on ground wearing just any pair of shoes? How would you discharge her? What are the consequences if she simply walks away?

Problems & Exercises

Exercise:

Problem:

Sketch the electric field lines in the vicinity of the conductor in [link] given the field was originally uniform and parallel to the object’s long axis. Is the resulting field small near the long side of the object?

Exercise:

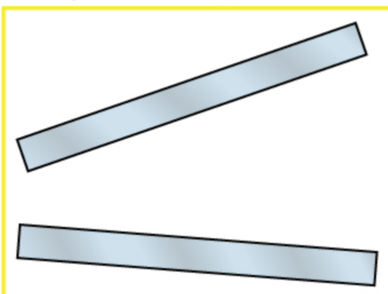
Problem:

Sketch the electric field lines in the vicinity of the conductor in [link] given the field was originally uniform and parallel to the object’s long axis. Is the resulting field small near the long side of the object?

Exercise:

Problem:

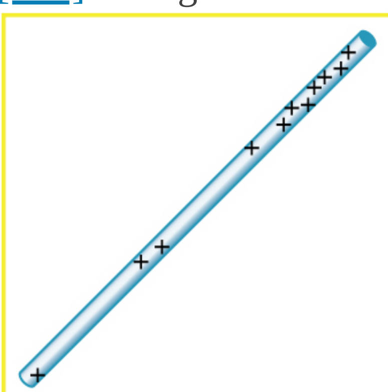
Sketch the electric field between the two conducting plates shown in [link], given the top plate is positive and an equal amount of negative charge is on the bottom plate. Be certain to indicate the distribution of charge on the plates.



Exercise:

Problem:

Sketch the electric field lines in the vicinity of the charged insulator in [link] noting its nonuniform charge distribution.



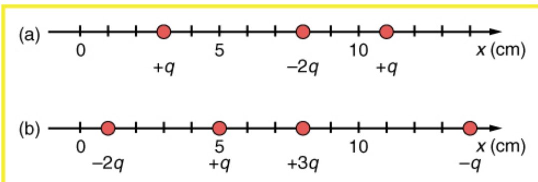
A charged insulating rod such as might be used in

a classroom demonstration.

Exercise:

Problem:

What is the force on the charge located at $x = 8.00\text{ cm}$ in [link](a) given that $q = 1.00\ \mu\text{C}$?



(a) Point charges located at 3.00, 8.00, and 11.0 cm along the *x*-axis. (b) Point charges located at 1.00, 5.00, 8.00, and 14.0 cm along the *x*-axis.

Exercise:

Problem:

(a) Find the total electric field at $x = 1.00\text{ cm}$ in [link](b) given that $q = 5.00\text{ nC}$. (b) Find the total electric field at $x = 11.00\text{ cm}$ in [link](b). (c) If the charges are allowed to move and eventually be brought to rest by friction, what will the final charge configuration be? (That is, will there be a single charge, double charge, etc., and what will its value(s) be?)

Solution:

(a) $E_{x=1.00\text{ cm}} = -\infty$

(b) $2.12 \times 10^5\ \mathrm{N/C}$

(c) one charge of $+q$

Exercise:

Problem:

(a) Find the electric field at $x = 5.00\ \mathrm{cm}$ in [link](a), given that $q = 1.00\ \mathrm{\mu C}$. (b) At what position between 3.00 and 8.00 cm is the total electric field the same as that for –$2q$ alone? (c) Can the electric field be zero anywhere between 0.00 and 8.00 cm? (d) At very large positive or negative values of *x*, the electric field approaches zero in both (a) and (b). In which does it most rapidly approach zero and why? (e) At what position to the right of 11.0 cm is the total electric field zero, other than at infinity? (Hint: A graphing calculator can yield considerable insight in this problem.)

Exercise:

Problem:

(a) Find the total Coulomb force on a charge of 2.00 nC located at $x = 4.00\ \mathrm{cm}$ in [link] (b), given that $q = 1.00\ \mathrm{\mu C}$. (b) Find the *x*-position at which the electric field is zero in [link] (b).

Solution:

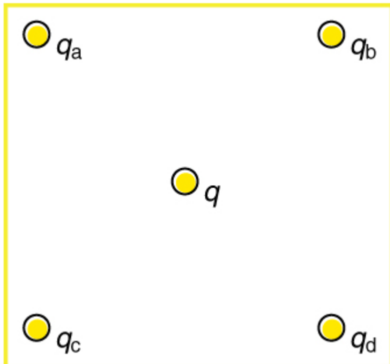
(a) 0.252 N to the left

(b) $x = 6.07\ \mathrm{cm}$

Exercise:

Problem:

Using the symmetry of the arrangement, determine the direction of the force on q in the figure below, given that $q_a = q_b$$=+7.50\ \mathrm{\mu C}$ and $q_c = q_d = -7.50\ \mathrm{\mu C}$. (b) Calculate the magnitude of the force on the charge q, given that the square is 10.0 cm on a side and $q = 2.00\ \mathrm{\mu C}$.



Exercise:

Problem:

(a) Using the symmetry of the arrangement, determine the direction of the electric field at the center of the square in [link], given that $q_a = q_b = -1.00\ \mu\text{C}$ and $q_c = q_d$=$+1.00\ \mu\text{C}$. (b) Calculate the magnitude of the electric field at the location of q, given that the square is 5.00 cm on a side.

Solution:

(a)The electric field at the center of the square will be straight up, since q_a and q_b are positive and q_c and q_d are negative and all have the same magnitude.

(b) $2.04 \times 10^7\ \text{N/C (upward)}$

Exercise:

Problem:

Find the electric field at the location of q_a in [link] given that $q_b = q_c = q_d$=$+2.00\ \text{nC}$, $q = -1.00\ \text{nC}$, and the square is 20.0 cm on a side.

Exercise:

Problem:

Find the total Coulomb force on the charge q in [link], given that $q = 1.00\ \mu\text{C}$, $q_a = 2.00\ \mu\text{C}$, $q_b = -3.00\ \mu\text{C}$, $q_c = -4.00\ \mu\text{C}$, and q_d=$+1.00\ \mu\text{C}$. The square is 50.0 cm on a side.

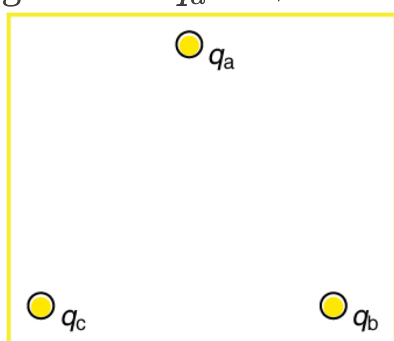
Solution:

$0.102\ \text{N}$ in the $-y$ direction

Exercise:

Problem:

(a) Find the electric field at the location of q_a in [link], given that $q_\text{b} = +10.00\ \mu\text{C}$ and $q_\text{c} = -5.00\ \mu\text{C}$. (b) What is the force on q_a, given that $q_\text{a} = +1.50\ \text{nC}$?



Point charges located at the corners of an equilateral triangle 25.0 cm on a side.

Exercise:

Problem:

(a) Find the electric field at the center of the triangular configuration of charges in [link], given that q_a=$+2.50\text{ nC}$, $q_b = -8.00\text{ nC}$, and q_c=$+1.50\text{ nC}$. (b) Is there any combination of charges, other than $q_a = q_b = q_c$, that will produce a zero strength electric field at the center of the triangular configuration?

Solution:

(a) $E = 4.36 \times 10^3\text{ N/C}\ \ 35.0$º, below the horizontal.

(b) No

Glossary

conductor
: an object with properties that allow charges to move about freely within it

free charge
: an electrical charge (either positive or negative) which can move about separately from its base molecule

electrostatic equilibrium
: an electrostatically balanced state in which all free electrical charges have stopped moving about

polarized
: a state in which the positive and negative charges within an object have collected in separate locations

ionosphere
: a layer of charged particles located around 100 km above the surface of Earth, which is responsible for a range of phenomena including the electric field surrounding Earth

Faraday cage

a metal shield which prevents electric charge from penetrating its surface

Applications of Electrostatics

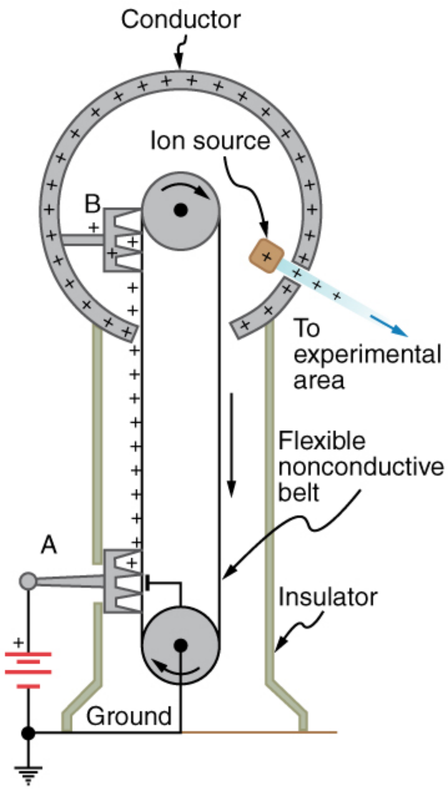
- Name several real-world applications of the study of electrostatics.

The study of **electrostatics** has proven useful in many areas. This module covers just a few of the many applications of electrostatics.

The Van de Graaff Generator

Van de Graaff generators (or Van de Graaffs) are not only spectacular devices used to demonstrate high voltage due to static electricity—they are also used for serious research. The first was built by Robert Van de Graaff in 1931 (based on original suggestions by Lord Kelvin) for use in nuclear physics research. [link] shows a schematic of a large research version. Van de Graaffs utilize both smooth and pointed surfaces, and conductors and insulators to generate large static charges and, hence, large voltages.

A very large excess charge can be deposited on the sphere, because it moves quickly to the outer surface. Practical limits arise because the large electric fields polarize and eventually ionize surrounding materials, creating free charges that neutralize excess charge or allow it to escape. Nevertheless, voltages of 15 million volts are well within practical limits.



Schematic of Van de Graaff generator. A battery (A) supplies excess positive charge to a pointed conductor, the points of which spray the charge onto a moving insulating belt near the bottom. The pointed conductor (B) on top in the large sphere picks up the charge. (The induced electric field at the points is so large that it removes the charge from the belt.) This can be done because the charge does not

remain inside the conducting sphere but moves to its outside surface. An ion source inside the sphere produces positive ions, which are accelerated away from the positive sphere to high velocities.

Note:
Take-Home Experiment: Electrostatics and Humidity
Rub a comb through your hair and use it to lift pieces of paper. It may help to tear the pieces of paper rather than cut them neatly. Repeat the exercise in your bathroom after you have had a long shower and the air in the bathroom is moist. Is it easier to get electrostatic effects in dry or moist air? Why would torn paper be more attractive to the comb than cut paper? Explain your observations.

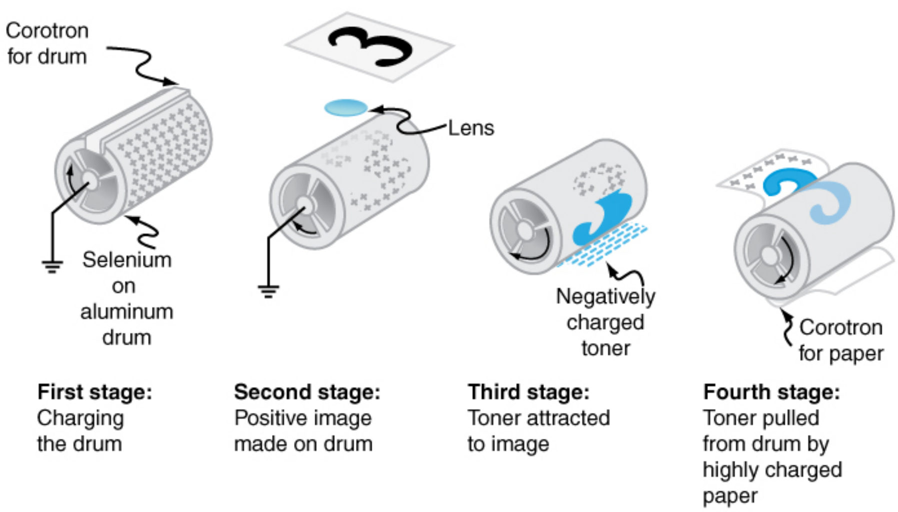
Xerography

Most copy machines use an electrostatic process called **xerography**—a word coined from the Greek words *xeros* for dry and *graphos* for writing. The heart of the process is shown in simplified form in [link].

A selenium-coated aluminum drum is sprayed with positive charge from points on a device called a corotron. Selenium is a substance with an interesting property—it is a **photoconductor**. That is, selenium is an insulator when in the dark and a conductor when exposed to light.

In the first stage of the xerography process, the conducting aluminum drum is **grounded** so that a negative charge is induced under the thin layer of uniformly positively charged selenium. In the second stage, the surface of the drum is exposed to the image of whatever is to be copied. Where the image is light, the selenium becomes conducting, and the positive charge is neutralized. In dark areas, the positive charge remains, and so the image has been transferred to the drum.

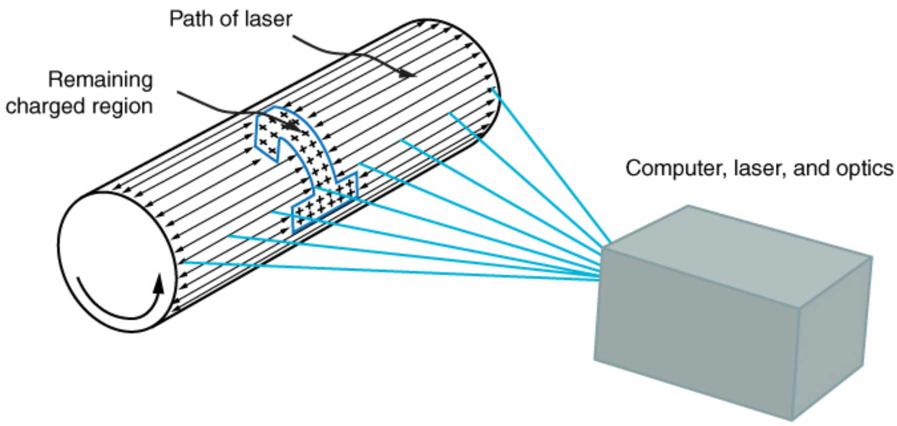
The third stage takes a dry black powder, called toner, and sprays it with a negative charge so that it will be attracted to the positive regions of the drum. Next, a blank piece of paper is given a greater positive charge than on the drum so that it will pull the toner from the drum. Finally, the paper and electrostatically held toner are passed through heated pressure rollers, which melt and permanently adhere the toner within the fibers of the paper.



Xerography is a dry copying process based on electrostatics. The major steps in the process are the charging of the photoconducting drum, transfer of an image creating a positive charge duplicate, attraction of toner to the charged parts of the drum, and transfer of toner to the paper. Not shown are heat treatment of the paper and cleansing of the drum for the next copy.

Laser Printers

Laser printers use the xerographic process to make high-quality images on paper, employing a laser to produce an image on the photoconducting drum as shown in [link]. In its most common application, the laser printer receives output from a computer, and it can achieve high-quality output because of the precision with which laser light can be controlled. Many laser printers do significant information processing, such as making sophisticated letters or fonts, and may contain a computer more powerful than the one giving them the raw data to be printed.

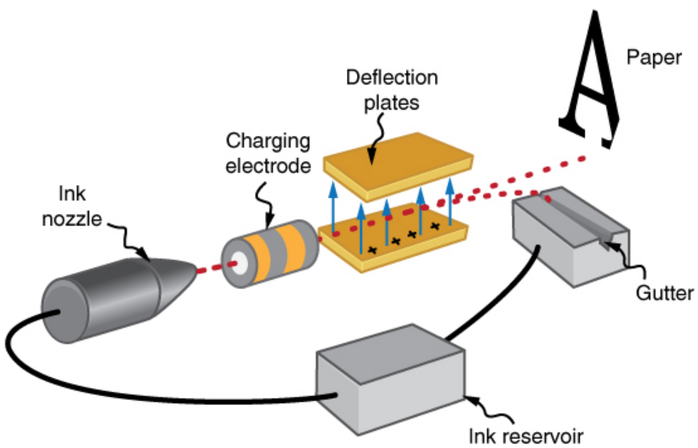


In a laser printer, a laser beam is scanned across a photoconducting drum, leaving a positive charge image. The other steps for charging the drum and transferring the image to paper are the same as in xerography. Laser light can be very precisely controlled, enabling laser printers to produce high-quality images.

Ink Jet Printers and Electrostatic Painting

The **ink jet printer**, commonly used to print computer-generated text and graphics, also employs electrostatics. A nozzle makes a fine spray of tiny ink droplets, which are then given an electrostatic charge. (See [link].)

Once charged, the droplets can be directed, using pairs of charged plates, with great precision to form letters and images on paper. Ink jet printers can produce color images by using a black jet and three other jets with primary colors, usually cyan, magenta, and yellow, much as a color television produces color. (This is more difficult with xerography, requiring multiple drums and toners.)



The nozzle of an ink-jet printer produces small ink droplets, which are sprayed with electrostatic charge. Various computer-driven devices are then used to direct the droplets to the correct positions on a page.

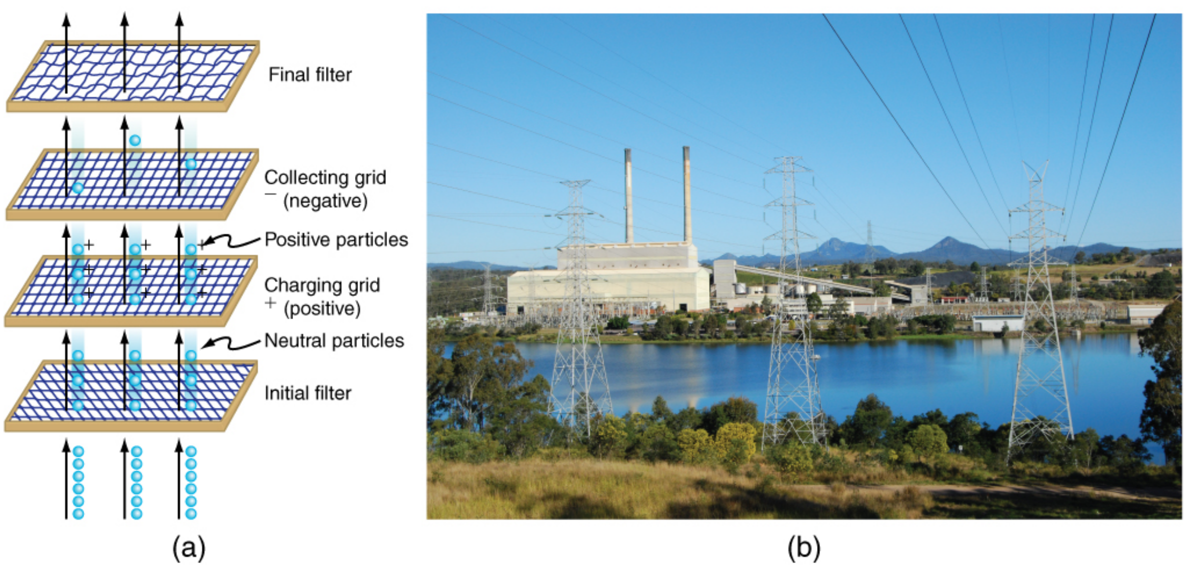
Electrostatic painting employs electrostatic charge to spray paint onto odd-shaped surfaces. Mutual repulsion of like charges causes the paint to fly away from its source. Surface tension forms drops, which are then attracted by unlike charges to the surface to be painted. Electrostatic painting can reach those hard-to-get at places, applying an even coat in a controlled

manner. If the object is a conductor, the electric field is perpendicular to the surface, tending to bring the drops in perpendicularly. Corners and points on conductors will receive extra paint. Felt can similarly be applied.

Smoke Precipitators and Electrostatic Air Cleaning

Another important application of electrostatics is found in air cleaners, both large and small. The electrostatic part of the process places excess (usually positive) charge on smoke, dust, pollen, and other particles in the air and then passes the air through an oppositely charged grid that attracts and retains the charged particles. (See [link].)

Large **electrostatic precipitators** are used industrially to remove over 99% of the particles from stack gas emissions associated with the burning of coal and oil. Home precipitators, often in conjunction with the home heating and air conditioning system, are very effective in removing polluting particles, irritants, and allergens.



(a) Schematic of an electrostatic precipitator. Air is passed through grids of opposite charge. The first grid charges airborne particles, while the second attracts and collects them. (b) The dramatic effect of

electrostatic precipitators is seen by the absence of smoke from this power plant. (credit: Cmdalgleish, Wikimedia Commons)

Note:
Problem-Solving Strategies for Electrostatics

1. Examine the situation to determine if static electricity is involved. This may concern separated stationary charges, the forces among them, and the electric fields they create.
2. Identify the system of interest. This includes noting the number, locations, and types of charges involved.
3. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful. Determine whether the Coulomb force is to be considered directly—if so, it may be useful to draw a free-body diagram, using electric field lines.
4. Make a list of what is given or can be inferred from the problem as stated (identify the knowns). It is important to distinguish the Coulomb force from the electric field , for example.
5. Solve the appropriate equation for the quantity to be determined (the unknown) or draw the field lines as requested.
6. Examine the answer to see if it is reasonable: Does it make sense? Are units correct and the numbers involved reasonable?

Integrated Concepts

The Integrated Concepts exercises for this module involve concepts such as electric charges, electric fields, and several other topics. Physics is most interesting when applied to general situations involving more than a narrow set of physical principles. The electric field exerts force on charges, for example, and hence the relevance of Dynamics: Force and Newton’s Laws of Motion. The following topics are involved in some or all of the problems labeled “Integrated Concepts”:

- Kinematics
- Two-Dimensional Kinematics
- Dynamics: Force and Newton’s Laws of Motion
- Uniform Circular Motion and Gravitation
- Statics and Torque
- Fluid Statics

The following worked example illustrates how this strategy is applied to an Integrated Concept problem:

Example:

Acceleration of a Charged Drop of Gasoline

If steps are not taken to ground a gasoline pump, static electricity can be placed on gasoline when filling your car’s tank. Suppose a tiny drop of gasoline has a mass of and is given a positive charge of . (a) Find the weight of the drop. (b) Calculate the electric force on the drop if there is an upward electric field of strength due to other static electricity in the vicinity. (c) Calculate the drop’s acceleration.

Strategy

To solve an integrated concept problem, we must first identify the physical principles involved and identify the chapters in which they are found. Part (a) of this example asks for weight. This is a topic of dynamics and is defined in Dynamics: Force and Newton’s Laws of Motion. Part (b) deals with electric force on a charge, a topic of Electric Charge and Electric Field. Part (c) asks for acceleration, knowing forces and mass. These are part of Newton’s laws, also found in Dynamics: Force and Newton’s Laws of Motion.

The following solutions to each part of the example illustrate how the specific problem-solving strategies are applied. These involve identifying knowns and unknowns, checking to see if the answer is reasonable, and so on.

Solution for (a)

Weight is mass times the acceleration due to gravity, as first expressed in

Equation:

Entering the given mass and the average acceleration due to gravity yields
Equation:

Discussion for (a)
This is a small weight, consistent with the small mass of the drop.
Solution for (b)
The force an electric field exerts on a charge is given by rearranging the following equation:
Equation:

Here we are given the charge (is twice the fundamental unit of charge) and the electric field strength, and so the electric force is found to be
Equation:

Discussion for (b)
While this is a small force, it is greater than the weight of the drop.
Solution for (c)
The acceleration can be found using Newton’s second law, provided we can identify all of the external forces acting on the drop. We assume only the drop’s weight and the electric force are significant. Since the drop has a positive charge and the electric field is given to be upward, the electric force is upward. We thus have a one-dimensional (vertical direction) problem, and we can state Newton’s second law as
Equation:

where . Entering this and the known values into the expression for Newton’s second law yields

Equation:

Discussion for (c)

This is an upward acceleration great enough to carry the drop to places where you might not wish to have gasoline.

This worked example illustrates how to apply problem-solving strategies to situations that include topics in different chapters. The first step is to identify the physical principles involved in the problem. The second step is to solve for the unknown using familiar problem-solving strategies. These are found throughout the text, and many worked examples show how to use them for single topics. In this integrated concepts example, you can see how to apply them across several topics. You will find these techniques useful in applications of physics outside a physics course, such as in your profession, in other science disciplines, and in everyday life. The following problems will build your skills in the broad application of physical principles.

Note:

Unreasonable Results

The Unreasonable Results exercises for this module have results that are unreasonable because some premise is unreasonable or because certain of the premises are inconsistent with one another. Physical principles applied correctly then produce unreasonable results. The purpose of these problems is to give practice in assessing whether nature is being accurately described, and if it is not to trace the source of difficulty.

Note:
Problem-Solving Strategy
To determine if an answer is reasonable, and to determine the cause if it is not, do the following.

1. Solve the problem using strategies as outlined above. Use the format followed in the worked examples in the text to solve the problem as usual.
2. Check to see if the answer is reasonable. Is it too large or too small, or does it have the wrong sign, improper units, and so on?
3. If the answer is unreasonable, look for what specifically could cause the identified difficulty. Usually, the manner in which the answer is unreasonable is an indication of the difficulty. For example, an extremely large Coulomb force could be due to the assumption of an excessively large separated charge.

Section Summary

- Electrostatics is the study of electric fields in static equilibrium.
- In addition to research using equipment such as a Van de Graaff generator, many practical applications of electrostatics exist, including photocopiers, laser printers, ink-jet printers and electrostatic air filters.

Problems & Exercises

Exercise:

Problem:

(a) What is the electric field 5.00 m from the center of the terminal of a Van de Graaff with a 3.00 mC charge, noting that the field is equivalent to that of a point charge at the center of the terminal? (b) At this distance, what force does the field exert on a charge on the Van de Graaff's belt?

Exercise:

Problem:

(a) What is the direction and magnitude of an electric field that supports the weight of a free electron near the surface of Earth? (b) Discuss what the small value for this field implies regarding the relative strength of the gravitational and electrostatic forces.

Solution:

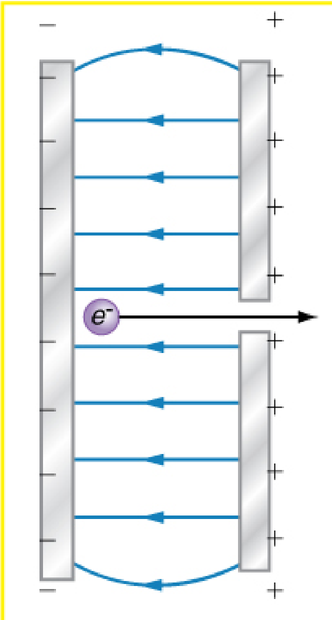
(a)

(b)the coulomb force is extraordinarily stronger than gravity

Exercise:

Problem:

A simple and common technique for accelerating electrons is shown in [link], where there is a uniform electric field between two plates. Electrons are released, usually from a hot filament, near the negative plate, and there is a small hole in the positive plate that allows the electrons to continue moving. (a) Calculate the acceleration of the electron if the field strength is . (b) Explain why the electron will not be pulled back to the positive plate once it moves through the hole.



Parallel conducting plates with opposite charges on them create a relatively uniform electric field used to accelerate electrons to the right. Those that go through the hole can be used to make a TV or computer screen glow or to produce X-rays.

Exercise:

Problem:

Earth has a net charge that produces an electric field of approximately 150 N/C downward at its surface. (a) What is the magnitude and sign of the excess charge, noting the electric field of a conducting sphere is equivalent to a point charge at its center? (b) What acceleration will the field produce on a free electron near Earth's surface? (c) What mass object with a single extra electron will have its weight supported by this field?

Solution:

(a)

(b)

(c)

Exercise:

Problem:

Point charges of and are placed 0.500 m apart. (a) At what point along the line between them is the electric field zero? (b) What is the electric field halfway between them?

Exercise:

Problem:

What can you say about two charges and , if the electric field one-fourth of the way from to is zero?

Solution:

The charge is 9 times greater than .

Exercise:

Problem: **Integrated Concepts**

Calculate the angular velocity ω of an electron orbiting a proton in the hydrogen atom, given the radius of the orbit is . You may assume that the proton is stationary and the centripetal force is supplied by Coulomb attraction.

Exercise:

Problem: Integrated Concepts

An electron has an initial velocity of in a uniform strength electric field. The field accelerates the electron in the direction opposite to its initial velocity. (a) What is the direction of the electric field? (b) How far does the electron travel before coming to rest? (c) How long does it take the electron to come to rest? (d) What is the electron's velocity when it returns to its starting point?

Exercise:

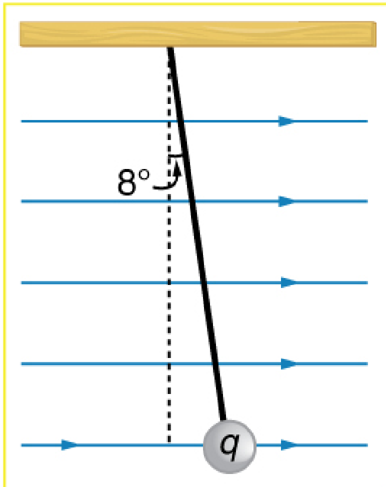
Problem: Integrated Concepts

The practical limit to an electric field in air is about . Above this strength, sparking takes place because air begins to ionize and charges flow, reducing the field. (a) Calculate the distance a free proton must travel in this field to reach of the speed of light, starting from rest. (b) Is this practical in air, or must it occur in a vacuum?

Exercise:

Problem: Integrated Concepts

A 5.00 g charged insulating ball hangs on a 30.0 cm long string in a uniform horizontal electric field as shown in [link]. Given the charge on the ball is , find the strength of the field.

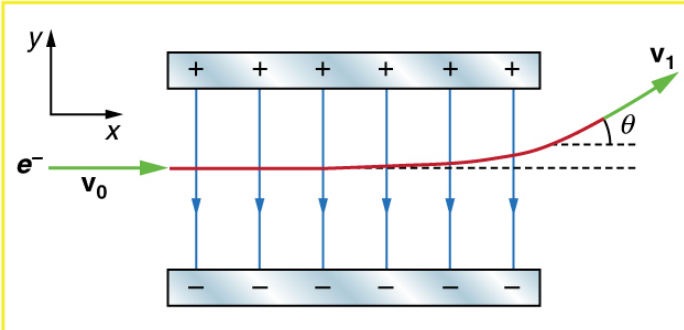


A horizontal electric field causes the charged ball to hang at an angle of °.

Exercise:

Problem: Integrated Concepts

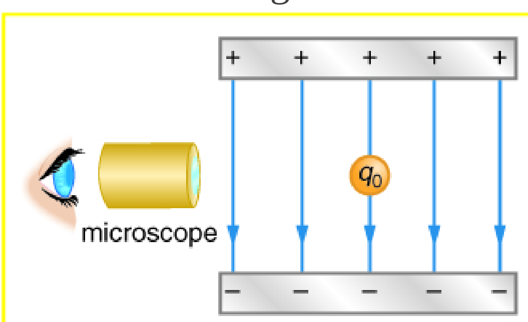
[link] shows an electron passing between two charged metal plates that create an 100 N/C vertical electric field perpendicular to the electron's original horizontal velocity. (These can be used to change the electron's direction, such as in an oscilloscope.) The initial speed of the electron is , and the horizontal distance it travels in the uniform field is 4.00 cm. (a) What is its vertical deflection? (b) What is the vertical component of its final velocity? (c) At what angle does it exit? Neglect any edge effects.



Exercise:

Problem: Integrated Concepts

The classic Millikan oil drop experiment was the first to obtain an accurate measurement of the charge on an electron. In it, oil drops were suspended against the gravitational force by a vertical electric field. (See [link].) Given the oil drop to be in radius and have a density of : (a) Find the weight of the drop. (b) If the drop has a single excess electron, find the electric field strength needed to balance its weight.



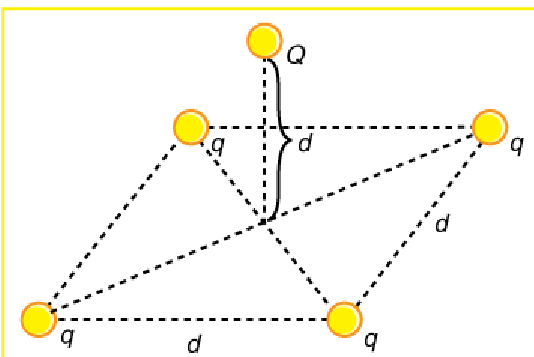
In the Millikan oil drop experiment, small drops can be suspended in an electric field by the force exerted on a single excess electron. Classically, this experiment was used to determine the electron charge by

measuring the electric field and mass of the drop.

Exercise:

Problem: Integrated Concepts

(a) In [link], four equal charges lie on the corners of a square. A fifth charge is on a mass directly above the center of the square, at a height equal to the length of one side of the square. Determine the magnitude of in terms of , , and , if the Coulomb force is to equal the weight of . (b) Is this equilibrium stable or unstable? Discuss.



Four equal charges on the corners of a horizontal square support the weight of a fifth charge located directly above the center of the square.

Exercise:

Problem: Unreasonable Results

(a) Calculate the electric field strength near a 10.0 cm diameter conducting sphere that has 1.00 C of excess charge on it. (b) What is

unreasonable about this result? (c) Which assumptions are responsible?

Exercise:

Problem: **Unreasonable Results**

(a) Two 0.500 g raindrops in a thunderhead are 1.00 cm apart when they each acquire 1.00 mC charges. Find their acceleration. (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?

Exercise:

Problem: **Unreasonable Results**

A wrecking yard inventor wants to pick up cars by charging a 0.400 m diameter ball and inducing an equal and opposite charge on the car. If a car has a 1000 kg mass and the ball is to be able to lift it from a distance of 1.00 m: (a) What minimum charge must be used? (b) What is the electric field near the surface of the ball? (c) Why are these results unreasonable? (d) Which premise or assumption is responsible?

Exercise:

Problem: **Construct Your Own Problem**

Consider two insulating balls with evenly distributed equal and opposite charges on their surfaces, held with a certain distance between the centers of the balls. Construct a problem in which you calculate the electric field (magnitude and direction) due to the balls at various points along a line running through the centers of the balls and extending to infinity on either side. Choose interesting points and comment on the meaning of the field at those points. For example, at what points might the field be just that due to one ball and where does the field become negligibly small? Among the things to be considered are the magnitudes of the charges and the distance between the centers of the balls. Your instructor may wish for you to consider the electric

field off axis or for a more complex array of charges, such as those in a water molecule.

Exercise:

Problem: Construct Your Own Problem

Consider identical spherical conducting space ships in deep space where gravitational fields from other bodies are negligible compared to the gravitational attraction between the ships. Construct a problem in which you place identical excess charges on the space ships to exactly counter their gravitational attraction. Calculate the amount of excess charge needed. Examine whether that charge depends on the distance between the centers of the ships, the masses of the ships, or any other factors. Discuss whether this would be an easy, difficult, or even impossible thing to do in practice.

Glossary

Van de Graaff generator
: a machine that produces a large amount of excess charge, used for experiments with high voltage

electrostatics
: the study of electric forces that are static or slow-moving

photoconductor
: a substance that is an insulator until it is exposed to light, when it becomes a conductor

xerography
: a dry copying process based on electrostatics

grounded
: connected to the ground with a conductor, so that charge flows freely to and from the Earth to the grounded object

laser printer

uses a laser to create a photoconductive image on a drum, which attracts dry ink particles that are then rolled onto a sheet of paper to print a high-quality copy of the image

ink-jet printer

small ink droplets sprayed with an electric charge are controlled by electrostatic plates to create images on paper

electrostatic precipitators

filters that apply charges to particles in the air, then attract those charges to a filter, removing them from the airstream

# Introduction to Electric Potential and Electric Energy
class="introduction"

Automated external defibrillator unit (AED) (credit: U.S. Defense Department photo/Tech. Sgt. Suzanne M. Day)

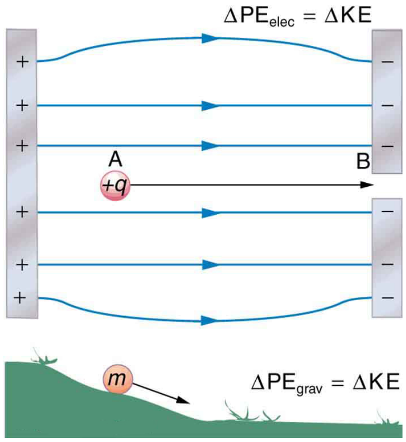
In Electric Charge and Electric Field, we just scratched the surface (or at least rubbed it) of electrical phenomena. Two of the most familiar aspects of

electricity are its energy and *voltage*. We know, for example, that great amounts of electrical energy can be stored in batteries, are transmitted cross-country through power lines, and may jump from clouds to explode the sap of trees. In a similar manner, at molecular levels, *ions* cross cell membranes and transfer information. We also know about voltages associated with electricity. Batteries are typically a few volts, the outlets in your home produce 120 volts, and power lines can be as high as hundreds of thousands of volts. But energy and voltage are not the same thing. A motorcycle battery, for example, is small and would not be very successful in replacing the much larger car battery, yet each has the same voltage. In this chapter, we shall examine the relationship between voltage and electrical energy and begin to explore some of the many applications of electricity.

Electric Potential Energy: Potential Difference

- Define electric potential and electric potential energy.
- Describe the relationship between potential difference and electrical potential energy.
- Explain electron volt and its usage in submicroscopic process.
- Determine electric potential energy given potential difference and amount of charge.

When a free positive charge *q* is accelerated by an electric field, such as shown in [link], it is given kinetic energy. The process is analogous to an object being accelerated by a gravitational field. It is as if the charge is going down an electrical hill where its electric potential energy is converted to kinetic energy. Let us explore the work done on a charge *q* by the electric field in this process, so that we may develop a definition of electric potential energy.



A charge accelerated by an electric field is analogous to a mass going down a hill. In both cases potential energy is converted to another form. Work is done by a force, but since this force

is conservative, we
can write
$W = -\Delta\text{PE}.$

The electrostatic or Coulomb force is conservative, which means that the work done on q is independent of the path taken. This is exactly analogous to the gravitational force in the absence of dissipative forces such as friction. When a force is conservative, it is possible to define a potential energy associated with the force, and it is usually easier to deal with the potential energy (because it depends only on position) than to calculate the work directly.

We use the letters PE to denote electric potential energy, which has units of joules (J). The change in potential energy, ΔPE, is crucial, since the work done by a conservative force is the negative of the change in potential energy; that is, $W = -\Delta\text{PE}$. For example, work W done to accelerate a positive charge from rest is positive and results from a loss in PE, or a negative ΔPE. There must be a minus sign in front of ΔPE to make W positive. PE can be found at any point by taking one point as a reference and calculating the work needed to move a charge to the other point.

Note:
Potential Energy
$W = -\Delta\text{PE}$. For example, work W done to accelerate a positive charge from rest is positive and results from a loss in PE, or a negative ΔPE. There must be a minus sign in front of ΔPE to make W positive. PE can be found at any point by taking one point as a reference and calculating the work needed to move a charge to the other point.

Gravitational potential energy and electric potential energy are quite analogous. Potential energy accounts for work done by a conservative force and gives added insight regarding energy and energy transformation

without the necessity of dealing with the force directly. It is much more common, for example, to use the concept of voltage (related to electric potential energy) than to deal with the Coulomb force directly.

Calculating the work directly is generally difficult, since $W = \text{Fd}\ \text{cos}\ \theta$ and the direction and magnitude of F can be complex for multiple charges, for odd-shaped objects, and along arbitrary paths. But we do know that, since $F = \text{qE}$, the work, and hence ΔPE, is proportional to the test charge q. To have a physical quantity that is independent of test charge, we define **electric potential** V (or simply potential, since electric is understood) to be the potential energy per unit charge:

Equation:

$$V = \frac{\text{PE}}{q}.$$

Note:

Electric Potential

This is the electric potential energy per unit charge.

Equation:

$$V = \frac{\text{PE}}{q}$$

Since PE is proportional to q, the dependence on q cancels. Thus V does not depend on q. The change in potential energy ΔPE is crucial, and so we are concerned with the difference in potential or potential difference ΔV between two points, where

Equation:

$$\Delta V = V_{\text{B}} - V_{\text{A}} = \frac{\Delta\text{PE}}{q}.$$

The **potential difference** between points A and B, $V_{\text{B}} - V_{\text{A}}$, is thus defined to be the change in potential energy of a charge q moved from A to B, divided by the charge. Units of potential difference are joules per coulomb, given the name volt (V) after Alessandro Volta.

Equation:

$$1\ \text{V} = 1\ \frac{\text{J}}{\text{C}}$$

Note:

Potential Difference

The potential difference between points A and B, $V_{\text{B}} - V_{\text{A}}$, is defined to be the change in potential energy of a charge q moved from A to B, divided by the charge. Units of potential difference are joules per coulomb, given the name volt (V) after Alessandro Volta.

Equation:

$$1\ \text{V} = 1\ \frac{\text{J}}{\text{C}}$$

The familiar term **voltage** is the common name for potential difference. Keep in mind that whenever a voltage is quoted, it is understood to be the potential difference between two points. For example, every battery has two terminals, and its voltage is the potential difference between them. More fundamentally, the point you choose to be zero volts is arbitrary. This is analogous to the fact that gravitational potential energy has an arbitrary zero, such as sea level or perhaps a lecture hall floor.

In summary, the relationship between potential difference (or voltage) and electrical potential energy is given by

Equation:

$$\Delta V = \frac{\Delta \text{PE}}{q} \text{ and } \Delta \text{PE} = q\Delta V.$$

Note:
Potential Difference and Electrical Potential Energy
The relationship between potential difference (or voltage) and electrical potential energy is given by
Equation:

$$\Delta V = \frac{\Delta \text{PE}}{q} \text{ and } \Delta \text{PE} = q\Delta V.$$

The second equation is equivalent to the first.

Voltage is not the same as energy. Voltage is the energy per unit charge. Thus a motorcycle battery and a car battery can both have the same voltage (more precisely, the same potential difference between battery terminals), yet one stores much more energy than the other since $\Delta \text{PE} = q\Delta V$. The car battery can move more charge than the motorcycle battery, although both are 12 V batteries.

Example:
Calculating Energy
Suppose you have a 12.0 V motorcycle battery that can move 5000 C of charge, and a 12.0 V car battery that can move 60,000 C of charge. How much energy does each deliver? (Assume that the numerical value of each charge is accurate to three significant figures.)
Strategy
To say we have a 12.0 V battery means that its terminals have a 12.0 V potential difference. When such a battery moves charge, it puts the charge

through a potential difference of 12.0 V, and the charge is given a change in potential energy equal to $\Delta\text{PE} = q\Delta V$.
So to find the energy output, we multiply the charge moved by the potential difference.
Solution
For the motorcycle battery, $q = 5000\text{ C}$ and $\Delta V = 12.0\text{ V}$. The total energy delivered by the motorcycle battery is
Equation:

$$\begin{aligned}\Delta\text{PE}_{\text{cycle}} &= (5000\text{ C})(12.0\text{ V}) \\ &= (5000\text{ C})(12.0\text{ J/C}) \\ &= 6.00\times10^{4}\text{ J}.\end{aligned}$$

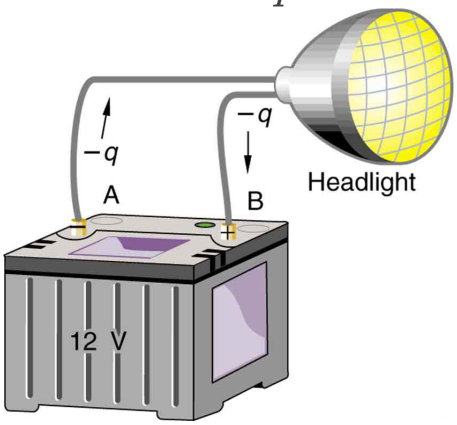
Similarly, for the car battery, $q = 60{,}000\text{ C}$ and
Equation:

$$\begin{aligned}\Delta\text{PE}_{\text{car}} &= (60{,}000\text{ C})(12.0\text{ V}) \\ &= 7.20\times10^{5}\text{ J}.\end{aligned}$$

Discussion
While voltage and energy are related, they are not the same thing. The voltages of the batteries are identical, but the energy supplied by each is quite different. Note also that as a battery is discharged, some of its energy is used internally and its terminal voltage drops, such as when headlights dim because of a low car battery. The energy supplied by the battery is still calculated as in this example, but not all of the energy is available for external use.

Note that the energies calculated in the previous example are absolute values. The change in potential energy for the battery is negative, since it loses energy. These batteries, like many electrical systems, actually move negative charge—electrons in particular. The batteries repel electrons from their negative terminals (A) through whatever circuitry is involved and attract them to their positive terminals (B) as shown in [link]. The change in potential is $\Delta V = V_{\text{B}} - V_{\text{A}} = +12\text{ V}$ and the charge q is negative, so that

$\Delta\text{PE} = q\Delta V$ is negative, meaning the potential energy of the battery has decreased when q has moved from A to B.



A battery moves negative charge from its negative terminal through a headlight to its positive terminal. Appropriate combinations of chemicals in the battery separate charges so that the negative terminal has an excess of negative charge, which is repelled by it and attracted to the excess positive charge on the other terminal. In terms of potential, the positive terminal is at a higher voltage than the negative. Inside the battery, both positive and negative charges move.

Example:
How Many Electrons Move through a Headlight Each Second?
When a 12.0 V car battery runs a single 30.0 W headlight, how many electrons pass through it each second?
Strategy
To find the number of electrons, we must first find the charge that moved in 1.00 s. The charge moved is related to voltage and energy through the equation $\Delta \mathrm{PE} = q\Delta V$. A 30.0 W lamp uses 30.0 joules per second. Since the battery loses energy, we have $\Delta \mathrm{PE} = –30.0\ \mathrm{J}$ and, since the electrons are going from the negative terminal to the positive, we see that $\Delta V = +12.0\ \mathrm{V}$.
Solution
To find the charge q moved, we solve the equation $\Delta \mathrm{PE} = q\Delta V$:
Equation:

$$q = \frac{\Delta \mathrm{PE}}{\Delta V}.$$

Entering the values for $\Delta \mathrm{PE}$ and ΔV, we get
Equation:

$$q = \frac{–30.0\ \mathrm{J}}{+12.0\ \mathrm{V}} = \frac{–30.0\ \mathrm{J}}{+12.0\ \mathrm{J/C}} = –2.50\ \mathrm{C}.$$

The number of electrons $\mathrm{n_e}$ is the total charge divided by the charge per electron. That is,
Equation:

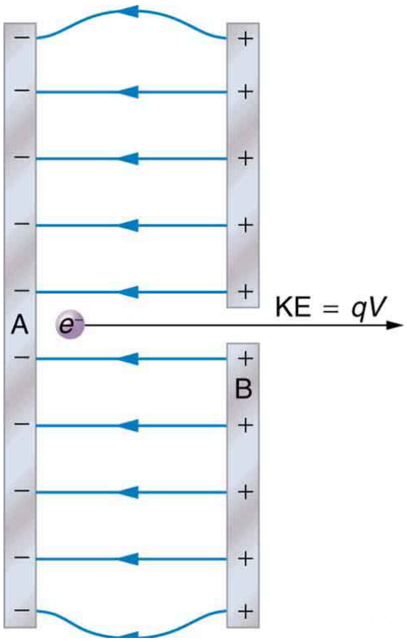
$$\mathrm{n_e} = \frac{–2.50\ \mathrm{C}}{–1.60\times10^{–19}\ \mathrm{C/e^–}} = 1.56\times10^{19}\ \text{electrons.}$$

Discussion
This is a very large number. It is no wonder that we do not ordinarily observe individual electrons with so many being present in ordinary

systems. In fact, electricity had been in use for many decades before it was determined that the moving charges in many circumstances were negative. Positive charge moving in the opposite direction of negative charge often produces identical effects; this makes it difficult to determine which is moving or whether both are moving.

The Electron Volt

The energy per electron is very small in macroscopic situations like that in the previous example—a tiny fraction of a joule. But on a submicroscopic scale, such energy per particle (electron, proton, or ion) can be of great importance. For example, even a tiny fraction of a joule can be great enough for these particles to destroy organic molecules and harm living tissue. The particle may do its damage by direct collision, or it may create harmful x rays, which can also inflict damage. It is useful to have an energy unit related to submicroscopic effects. [link] shows a situation related to the definition of such an energy unit. An electron is accelerated between two charged metal plates as it might be in an old-model television tube or oscilloscope. The electron is given kinetic energy that is later converted to another form—light in the television tube, for example. (Note that downhill for the electron is uphill for a positive charge.) Since energy is related to voltage by $\Delta \mathrm{PE} = q\Delta V$, we can think of the joule as a coulomb-volt.



A typical electron gun accelerates electrons using a potential difference between two metal plates. The energy of the electron in electron volts is numerically the same as the voltage between the plates. For example, a 5000 V potential difference produces 5000 eV electrons.

On the submicroscopic scale, it is more convenient to define an energy unit called the **electron volt** (eV), which is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,

Equation:

$$\begin{aligned} 1 \text{ eV} &= 1.60 \times 10^{-19} \text{ C } (1 \text{ V}) = 1.60 \times 10^{-19} \text{ C } (1 \text{ J/C}) \\ &= 1.60 \times 10^{-19} \text{ J.} \end{aligned}$$

Note:
Electron Volt
On the submicroscopic scale, it is more convenient to define an energy unit called the electron volt (eV), which is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,
Equation:

$$\begin{aligned} 1 \text{ eV} &= 1.60 \times 10^{-19} \text{ C } (1 \text{ V}) = 1.60 \times 10^{-19} \text{ C } (1 \text{ J/C}) \\ &= 1.60 \times 10^{-19} \text{ J.} \end{aligned}$$

An electron accelerated through a potential difference of 1 V is given an energy of 1 eV. It follows that an electron accelerated through 50 V is given 50 eV. A potential difference of 100,000 V (100 kV) will give an electron an energy of 100,000 eV (100 keV), and so on. Similarly, an ion with a double positive charge accelerated through 100 V will be given 200 eV of energy. These simple relationships between accelerating voltage and particle charges make the electron volt a simple and convenient energy unit in such circumstances.

Note:
Connections: Energy Units
The electron volt (eV) is the most common energy unit for submicroscopic processes. This will be particularly noticeable in the chapters on modern physics. Energy is so important to so many subjects that there is a tendency to define a special energy unit for each major topic. There are, for example,

calories for food energy, kilowatt-hours for electrical energy, and therms for natural gas energy.

The electron volt is commonly employed in submicroscopic processes—chemical valence energies and molecular and nuclear binding energies are among the quantities often expressed in electron volts. For example, about 5 eV of energy is required to break up certain organic molecules. If a proton is accelerated from rest through a potential difference of 30 kV, it is given an energy of 30 keV (30,000 eV) and it can break up as many as 6000 of these molecules ($30{,}000\text{ eV} \div 5\text{ eV per molecule} = 6000\text{ molecules}$). Nuclear decay energies are on the order of 1 MeV (1,000,000 eV) per event and can, thus, produce significant biological damage.

Conservation of Energy

The total energy of a system is conserved if there is no net addition (or subtraction) of work or heat transfer. For conservative forces, such as the electrostatic force, conservation of energy states that mechanical energy is a constant.

Mechanical energy is the sum of the kinetic energy and potential energy of a system; that is, $\text{KE} + \text{PE} = \text{constant}$. A loss of PE of a charged particle becomes an increase in its KE. Here PE is the electric potential energy. Conservation of energy is stated in equation form as
Equation:

$$\text{KE} + \text{PE} = \text{constant}$$

or
Equation:

$$\text{KE}_{\text{i}} + \text{PE}_{\text{i}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}},$$

where i and f stand for initial and final conditions. As we have found many times before, considering energy can give us insights and facilitate problem solving.

Example:
Electrical Potential Energy Converted to Kinetic Energy
Calculate the final speed of a free electron accelerated from rest through a potential difference of 100 V. (Assume that this numerical value is accurate to three significant figures.)
Strategy
We have a system with only conservative forces. Assuming the electron is accelerated in a vacuum, and neglecting the gravitational force (we will check on this assumption later), all of the electrical potential energy is converted into kinetic energy. We can identify the initial and final forms of energy to be $\mathrm{KE_i} = 0$, $\mathrm{KE_f} = ½ mv^2$, $\mathrm{PE_i} = qV$, and $\mathrm{PE_f} = 0$.
Solution
Conservation of energy states that
Equation:

$$\mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f}.$$

Entering the forms identified above, we obtain
Equation:

$$qV = \frac{mv^2}{2}.$$

We solve this for v:
Equation:

$$v = \sqrt{\frac{2\mathrm{qV}}{m}}.$$

Entering values for q, V, and m gives
Equation:

$$
\begin{aligned}
v &= \sqrt{\frac{2\left(-1.60\times10^{-19}\text{ C}\right)(-100\text{ J/C})}{9.11\times10^{-31}\text{ kg}}} \\
&= 5.93\times10^{6}\text{ m/s.}
\end{aligned}
$$

Discussion
Note that both the charge and the initial voltage are negative, as in [link]. From the discussions in Electric Charge and Electric Field, we know that electrostatic forces on small particles are generally very large compared with the gravitational force. The large final speed confirms that the gravitational force is indeed negligible here. The large speed also indicates how easy it is to accelerate electrons with small voltages because of their very small mass. Voltages much higher than the 100 V in this problem are typically used in electron guns. Those higher voltages produce electron speeds so great that relativistic effects must be taken into account. That is why a low voltage is considered (accurately) in this example.

Section Summary

- Electric potential is potential energy per unit charge.
- The potential difference between points A and B, $V_B - V_A$, defined to be the change in potential energy of a charge q moved from A to B, is equal to the change in potential energy divided by the charge, Potential difference is commonly called voltage, represented by the symbol ΔV.
  Equation:

$$\Delta V = \frac{\Delta \text{PE}}{q} \text{ and } \Delta \text{PE} = q\Delta V.$$

- An electron volt is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,
  Equation:

$$
\begin{aligned}
1\text{ eV} &= \left(1.60\times10^{-19}\text{ C}\right)(1\text{ V}) = \left(1.60\times10^{-19}\text{ C}\right)(1\text{ J/C}) \\
&= 1.60\times10^{-19}\text{ J.}
\end{aligned}
$$

- Mechanical energy is the sum of the kinetic energy and potential energy of a system, that is, $\mathrm{KE} + \mathrm{PE}$. This sum is a constant.

Conceptual Questions

Exercise:

Problem:

Voltage is the common word for potential difference. Which term is more descriptive, voltage or potential difference?

Exercise:

Problem:

If the voltage between two points is zero, can a test charge be moved between them with zero net work being done? Can this necessarily be done without exerting a force? Explain.

Exercise:

Problem:

What is the relationship between voltage and energy? More precisely, what is the relationship between potential difference and electric potential energy?

Exercise:

Problem: Voltages are always measured between two points. Why?

Exercise:

Problem:

How are units of volts and electron volts related? How do they differ?

Problems & Exercises

Exercise:

Problem:

Find the ratio of speeds of an electron and a negative hydrogen ion (one having an extra electron) accelerated through the same voltage, assuming non-relativistic final speeds. Take the mass of the hydrogen ion to be 1.67×10^{-27} kg.

Solution:

42.8

Exercise:

Problem:

An evacuated tube uses an accelerating voltage of 40 kV to accelerate electrons to hit a copper plate and produce x rays. Non-relativistically, what would be the maximum speed of these electrons?

Exercise:

Problem:

A bare helium nucleus has two positive charges and a mass of 6.64×10^{-27} kg. (a) Calculate its kinetic energy in joules at 2.00% of the speed of light. (b) What is this in electron volts? (c) What voltage would be needed to obtain this energy?

Exercise:

Problem: Integrated Concepts

Singly charged gas ions are accelerated from rest through a voltage of 13.0 V. At what temperature will the average kinetic energy of gas molecules be the same as that given these ions?

Solution:

1.00×10^{5} K

Exercise:

Problem: **Integrated Concepts**

The temperature near the center of the Sun is thought to be 15 million degrees Celsius $1.5 \times 10^7 \ ^\circ\text{C}$. Through what voltage must a singly charged ion be accelerated to have the same energy as the average kinetic energy of ions at this temperature?

Exercise:

Problem: **Integrated Concepts**

(a) What is the average power output of a heart defibrillator that dissipates 400 J of energy in 10.0 ms? (b) Considering the high-power output, why doesn't the defibrillator produce serious burns?

Solution:

(a) $4 \times 10^4 \text{ W}$

(b) A defibrillator does not cause serious burns because the skin conducts electricity well at high voltages, like those used in defibrillators. The gel used aids in the transfer of energy to the body, and the skin doesn't absorb the energy, but rather lets it pass through to the heart.

Exercise:

Problem: **Integrated Concepts**

A lightning bolt strikes a tree, moving 20.0 C of charge through a potential difference of 1.00×10^2 MV. (a) What energy was dissipated? (b) What mass of water could be raised from 15ºC to the boiling point and then boiled by this energy? (c) Discuss the damage that could be caused to the tree by the expansion of the boiling steam.

Exercise:

Problem: **Integrated Concepts**

A 12.0 V battery-operated bottle warmer heats 50.0 g of glass, 2.50×10^2 g of baby formula, and 2.00×10^2 g of aluminum from 20.0°C to 90.0°C. (a) How much charge is moved by the battery? (b) How many electrons per second flow if it takes 5.00 min to warm the formula? (Hint: Assume that the specific heat of baby formula is about the same as the specific heat of water.)

Solution:

(a) $7.40 \times 10^3\ \text{C}$

(b) 1.54×10^{20} electrons per second

Exercise:

Problem: **Integrated Concepts**

A battery-operated car utilizes a 12.0 V system. Find the charge the batteries must be able to move in order to accelerate the 750 kg car from rest to 25.0 m/s, make it climb a 2.00×10^2 m high hill, and then cause it to travel at a constant 25.0 m/s by exerting a 5.00×10^2 N force for an hour.

Solution:

$3.89 \times 10^6\ \text{C}$

Exercise:

Problem: **Integrated Concepts**

Fusion probability is greatly enhanced when appropriate nuclei are brought close together, but mutual Coulomb repulsion must be overcome. This can be done using the kinetic energy of high-temperature gas ions or by accelerating the nuclei toward one another.

(a) Calculate the potential energy of two singly charged nuclei separated by 1.00×10^{-12} m by finding the voltage of one at that distance and multiplying by the charge of the other. (b) At what temperature will atoms of a gas have an average kinetic energy equal to this needed electrical potential energy?

Exercise:

Problem: Unreasonable Results

(a) Find the voltage near a 10.0 cm diameter metal sphere that has 8.00 C of excess positive charge on it. (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Solution:

(a) 1.44×10^{12} V

(b) This voltage is very high. A 10.0 cm diameter sphere could never maintain this voltage; it would discharge.

(c) An 8.00 C charge is more charge than can reasonably be accumulated on a sphere of that size.

Exercise:

Problem: Construct Your Own Problem

Consider a battery used to supply energy to a cellular phone. Construct a problem in which you determine the energy that must be supplied by the battery, and then calculate the amount of charge it must be able to move in order to supply this energy. Among the things to be considered are the energy needs and battery voltage. You may need to look ahead to interpret manufacturer’s battery ratings in ampere-hours as energy in joules.

Glossary

electric potential
    potential energy per unit charge

potential difference (or voltage)
    change in potential energy of a charge moved from one point to another, divided by the charge; units of potential difference are joules per coulomb, known as volt

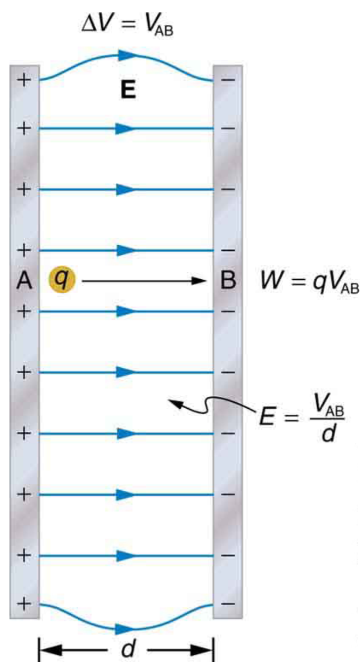
electron volt
    the energy given to a fundamental charge accelerated through a potential difference of one volt

mechanical energy
    sum of the kinetic energy and potential energy of a system; this sum is a constant

Electric Potential in a Uniform Electric Field

- Describe the relationship between voltage and electric field.
- Derive an expression for the electric potential and electric field.
- Calculate electric field strength given distance and voltage.

In the previous section, we explored the relationship between voltage and energy. In this section, we will explore the relationship between voltage and electric field. For example, a uniform electric field $\mathbf{E}$ is produced by placing a potential difference (or voltage) ΔV across two parallel metal plates, labeled A and B. (See [link].) Examining this will tell us what voltage is needed to produce a certain electric field strength; it will also reveal a more fundamental relationship between electric potential and electric field. From a physicist's point of view, either ΔV or $\mathbf{E}$ can be used to describe any charge distribution. ΔV is most closely tied to energy, whereas $\mathbf{E}$ is most closely related to force. ΔV is a **scalar** quantity and has no direction, while $\mathbf{E}$ is a **vector** quantity, having both magnitude and direction. (Note that the magnitude of the electric field strength, a scalar quantity, is represented by E below.) The relationship between ΔV and $\mathbf{E}$ is revealed by calculating the work done by the force in moving a charge from point A to point B. But, as noted in Electric Potential Energy: Potential Difference, this is complex for arbitrary charge distributions, requiring calculus. We therefore look at a uniform electric field as an interesting special case.



The relationship between V and E for parallel conducting plates is $E = V/d$. (Note that $\Delta V = V_{\text{AB}}$ in magnitude. For a charge that is moved from plate A at higher potential to plate B at lower potential, a minus sign needs to be included as follows: $-\Delta\ V = V_{\text{A}} - V_{\text{B}} = V_{\text{AB}}$. See the text for details.)

The work done by the electric field in [link] to move a positive charge q from A, the positive plate, higher potential, to B, the negative plate, lower potential, is

Equation:

$$W = -\Delta \text{PE} = -q\Delta V.$$

The potential difference between points A and B is
Equation:

$$-\Delta\ V = -(V_{\text{B}} - V_{\text{A}}) = V_{\text{A}} - V_{\text{B}} = V_{\text{AB}}.$$

Entering this into the expression for work yields
Equation:

$$W = qV_{\text{AB}}.$$

Work is $W = Fd\cos\theta$; here $\cos\theta = 1$, since the path is parallel to the field, and so $W = Fd$. Since $F = qE$, we see that $W = qEd$. Substituting this expression for work into the previous equation gives
Equation:

$$qEd = qV_{\text{AB}}.$$

The charge cancels, and so the voltage between points A and B is seen to be
Equation:

$$\left.\begin{aligned} V_{\text{AB}} &= Ed \\ E &= \frac{V_{\text{AB}}}{d} \end{aligned}\right\}(\text{uniform } E\text{ - field only}),$$

where d is the distance from A to B, or the distance between the plates in [link]. Note that the above equation implies the units for electric field are volts per meter. We already know the units for electric field are newtons per coulomb; thus the following relation among units is valid:
Equation:

$$1\ \text{N/C} = 1\ \text{V/m}.$$

Note:
Voltage between Points A and B
Equation:

$$\left.\begin{array}{r}V_{\text{AB}} = Ed \\ E = \frac{V_{\text{AB}}}{d}\end{array}\right\}(\text{uniform } E \text{ - field only}),$$

where d is the distance from A to B, or the distance between the plates.

Example:
What Is the Highest Voltage Possible between Two Plates?
Dry air will support a maximum electric field strength of about $3.0\times10^{6}\ \text{V/m}$. Above that value, the field creates enough ionization in the air to make the air a conductor. This allows a discharge or spark that reduces the field. What, then, is the maximum voltage between two parallel conducting plates separated by 2.5 cm of dry air?
Strategy
We are given the maximum electric field E between the plates and the distance d between them. The equation $V_{\text{AB}} = Ed$ can thus be used to calculate the maximum voltage.
Solution
The potential difference or voltage between the plates is
Equation:

$$\text{V}_{\text{AB}} = Ed.$$

Entering the given values for E and d gives
Equation:

$$V_{\text{AB}} = (3.0\times10^{6}\ \text{V/m})(0.025\ \text{m}) = 7.5\times10^{4}\ \text{V}$$

or
Equation:

$$V_{\text{AB}} = 75\ \text{kV}.$$

(The answer is quoted to only two digits, since the maximum field strength is approximate.)

Discussion

One of the implications of this result is that it takes about 75 kV to make a spark jump across a 2.5 cm (1 in.) gap, or 150 kV for a 5 cm spark. This limits the voltages that can exist between conductors, perhaps on a power transmission line. A smaller voltage will cause a spark if there are points on the surface, since points create greater fields than smooth surfaces. Humid air breaks down at a lower field strength, meaning that a smaller voltage will make a spark jump through humid air. The largest voltages can be built up, say with static electricity, on dry days.

A spark chamber is used to trace the paths of high-energy particles. Ionization created by the particles as they pass through the gas between the plates allows a spark to jump. The sparks are

perpendicular to the plates, following electric field lines between them. The potential difference between adjacent plates is not high enough to cause sparks without the ionization produced by particles from accelerator experiments (or cosmic rays). (credit: Daderot, Wikimedia Commons)

Example:
Field and Force inside an Electron Gun
(a) An electron gun has parallel plates separated by 4.00 cm and gives electrons 25.0 keV of energy. What is the electric field strength between the plates? (b) What force would this field exert on a piece of plastic with a $0.500\ \mu\text{C}$ charge that gets between the plates?
Strategy
Since the voltage and plate separation are given, the electric field strength can be calculated directly from the expression $E = \frac{V_{\text{AB}}}{d}$. Once the electric field strength is known, the force on a charge is found using $\mathbf{F} = q\ \mathbf{E}$. Since the electric field is in only one direction, we can write this equation in terms of the magnitudes, $F = q\ E$.
Solution for (a)
The expression for the magnitude of the electric field between two uniform metal plates is

Equation:

$$E = \frac{V_{\text{AB}}}{d}.$$

Since the electron is a single charge and is given 25.0 keV of energy, the potential difference must be 25.0 kV. Entering this value for V_{AB} and the plate separation of 0.0400 m, we obtain

Equation:

$$E = \frac{25.0\ \text{kV}}{0.0400\ \text{m}} = 6.25 \times 10^5\ \text{V/m}.$$

Solution for (b)

The magnitude of the force on a charge in an electric field is obtained from the equation

Equation:

$$F = qE.$$

Substituting known values gives

Equation:

$$F = (0.500 \times 10^{-6}\ \text{C})(6.25 \times 10^5\ \text{V/m}) = 0.313\ \text{N}.$$

Discussion

Note that the units are newtons, since $1\ \text{V/m} = 1\ \text{N/C}$. The force on the charge is the same no matter where the charge is located between the plates. This is because the electric field is uniform between the plates.

In more general situations, regardless of whether the electric field is uniform, it points in the direction of decreasing potential, because the force on a positive charge is in the direction of $\mathbf{E}$ and also in the direction of lower potential V. Furthermore, the magnitude of $\mathbf{E}$ equals the rate of decrease of V with distance. The faster V decreases over distance, the

greater the electric field. In equation form, the general relationship between voltage and electric field is

Equation:

$$E = -\frac{\Delta V}{\Delta s},$$

where Δs is the distance over which the change in potential, ΔV, takes place. The minus sign tells us that $\mathbf{E}$ points in the direction of decreasing potential. The electric field is said to be the *gradient* (as in grade or slope) of the electric potential.

Note:

Relationship between Voltage and Electric Field

In equation form, the general relationship between voltage and electric field is

Equation:

$$E = -\frac{\Delta V}{\Delta s},$$

where Δs is the distance over which the change in potential, ΔV, takes place. The minus sign tells us that $\mathbf{E}$ points in the direction of decreasing potential. The electric field is said to be the *gradient* (as in grade or slope) of the electric potential.

For continually changing potentials, ΔV and Δs become infinitesimals and differential calculus must be employed to determine the electric field.

Section Summary

- The voltage between points A and B is
  Equation:

$$\left.\begin{array}{l} V_{\mathrm{AB}} = Ed \\ E = \frac{V_{\mathrm{AB}}}{d} \end{array}\right\} (\text{uniform } E \text{ - field only}),$$

where d is the distance from A to B, or the distance between the plates.

- In equation form, the general relationship between voltage and electric field is

  Equation:

  $$E = -\frac{\Delta V}{\Delta s},$$

  where Δs is the distance over which the change in potential, ΔV, takes place. The minus sign tells us that $\mathbf{E}$ points in the direction of decreasing potential.) The electric field is said to be the *gradient* (as in grade or slope) of the electric potential.

Conceptual Questions

Exercise:

Problem:

Discuss how potential difference and electric field strength are related. Give an example.

Exercise:

Problem:

What is the strength of the electric field in a region where the electric potential is constant?

Exercise:

Problem:

Will a negative charge, initially at rest, move toward higher or lower potential? Explain why.

Problems & Exercises

Exercise:

Problem:

Show that units of V/m and N/C for electric field strength are indeed equivalent.

Exercise:

Problem:

What is the strength of the electric field between two parallel conducting plates separated by 1.00 cm and having a potential difference (voltage) between them of $1.50 \times 10^4 \text{ V}$?

Exercise:

Problem:

The electric field strength between two parallel conducting plates separated by 4.00 cm is $7.50 \times 10^4 \text{ V/m}$. (a) What is the potential difference between the plates? (b) The plate with the lowest potential is taken to be at zero volts. What is the potential 1.00 cm from that plate (and 3.00 cm from the other)?

Solution:

(a) 3.00 kV

(b) 750 V

Exercise:

Problem:

How far apart are two conducting plates that have an electric field strength of $4.50 \times 10^3 \text{ V/m}$ between them, if their potential difference is 15.0 kV?

Exercise:

Problem:

(a) Will the electric field strength between two parallel conducting plates exceed the breakdown strength for air ($3.0 \times 10^6 \text{ V/m}$) if the plates are separated by 2.00 mm and a potential difference of $5.0 \times 10^3 \text{ V}$ is applied? (b) How close together can the plates be with this applied voltage?

Solution:

(a) No. The electric field strength between the plates is $2.5 \times 10^6 \text{ V/m}$, which is lower than the breakdown strength for air ($3.0 \times 10^6 \text{ V/m}$).

(b) 1.7 mm

Exercise:

Problem:

The voltage across a membrane forming a cell wall is 80.0 mV and the membrane is 9.00 nm thick. What is the electric field strength? (The value is surprisingly large, but correct. Membranes are discussed in Capacitors and Dielectrics and Nerve Conduction—Electrocardiograms.) You may assume a uniform electric field.

Exercise:

Problem:

Membrane walls of living cells have surprisingly large electric fields across them due to separation of ions. (Membranes are discussed in some detail in Nerve Conduction—Electrocardiograms.) What is the voltage across an 8.00 nm–thick membrane if the electric field strength across it is 5.50 MV/m? You may assume a uniform electric field.

Solution:

44.0 mV

Exercise:

Problem:

Two parallel conducting plates are separated by 10.0 cm, and one of them is taken to be at zero volts. (a) What is the electric field strength between them, if the potential 8.00 cm from the zero volt plate (and 2.00 cm from the other) is 450 V? (b) What is the voltage between the plates?

Exercise:

Problem:

Find the maximum potential difference between two parallel conducting plates separated by 0.500 cm of air, given the maximum sustainable electric field strength in air to be $3.0 \times 10^6 \text{ V/m}$.

Solution:

15 kV

Exercise:

Problem:

A doubly charged ion is accelerated to an energy of 32.0 keV by the electric field between two parallel conducting plates separated by 2.00 cm. What is the electric field strength between the plates?

Exercise:

Problem:

An electron is to be accelerated in a uniform electric field having a strength of $2.00 \times 10^6 \text{ V/m}$. (a) What energy in keV is given to the electron if it is accelerated through 0.400 m? (b) Over what distance would it have to be accelerated to increase its energy by 50.0 GeV?

Solution:

(a) 800 KeV

(b) 25.0 km

Glossary

scalar
: physical quantity with magnitude but no direction

vector
: physical quantity with both magnitude and direction

Electrical Potential Due to a Point Charge

- Explain point charges and express the equation for electric potential of a point charge.
- Distinguish between electric potential and electric field.
- Determine the electric potential of a point charge given charge and distance.

Point charges, such as electrons, are among the fundamental building blocks of matter. Furthermore, spherical charge distributions (like on a metal sphere) create external electric fields exactly like a point charge. The electric potential due to a point charge is, thus, a case we need to consider. Using calculus to find the work needed to move a test charge q from a large distance away to a distance of r from a point charge Q, and noting the connection between work and potential $(W = -q\Delta V)$, it can be shown that the *electric potential V of a point charge* is

Equation:

$$V = \frac{kQ}{r} \text{ (Point Charge)},$$

where *k* is a constant equal to $9.0 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2$.

Note:

Electric Potential V of a Point Charge

The electric potential V of a point charge is given by

Equation:

$$V = \frac{kQ}{r} \text{ (Point Charge)}.$$

The potential at infinity is chosen to be zero. Thus V for a point charge decreases with distance, whereas $\mathbf{E}$ for a point charge decreases with distance squared:

Equation:

$$E = \frac{F}{q} = \frac{kQ}{r^2}.$$

Recall that the electric potential V is a scalar and has no direction, whereas the electric field $\mathbf{E}$ is a vector. To find the voltage due to a combination of point charges, you add the individual voltages as numbers. To find the total electric field, you must add the individual fields as *vectors*, taking magnitude and direction into account. This is consistent with the fact that V is closely associated with energy, a scalar, whereas $\mathbf{E}$ is closely associated with force, a vector.

Example:

What Voltage Is Produced by a Small Charge on a Metal Sphere?

Charges in static electricity are typically in the nanocoulomb (nC) to microcoulomb $(\mu\mathrm{C})$ range. What is the voltage 5.00 cm away from the center of a 1-cm diameter metal sphere that has a $-3.00\ \mathrm{nC}$ static charge?

Strategy

As we have discussed in Electric Charge and Electric Field, charge on a metal sphere spreads out uniformly and produces a field like that of a point charge located at its center. Thus we can find the voltage using the equation $V = \mathrm{kQ}/r$.

Solution

Entering known values into the expression for the potential of a point charge, we obtain

Equation:

$$\begin{aligned} V &= k\frac{Q}{r} \\ &= \left(8.99 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2\right)\left(\frac{-3.00\times10^{-9}\text{ C}}{5.00\times10^{-2}\text{ m}}\right) \\ &= -539 \text{ V}. \end{aligned}$$

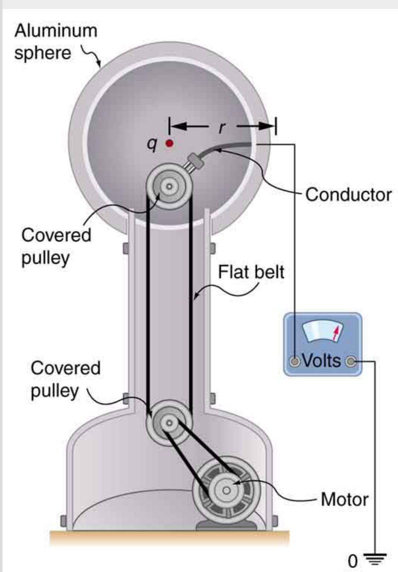
Discussion

The negative value for voltage means a positive charge would be attracted from a larger distance, since the potential is lower (more negative) than at larger distances. Conversely, a negative charge would be repelled, as expected.

Example:

What Is the Excess Charge on a Van de Graaff Generator

A demonstration Van de Graaff generator has a 25.0 cm diameter metal sphere that produces a voltage of 100 kV near its surface. (See [link].) What excess charge resides on the sphere? (Assume that each numerical value here is shown with three significant figures.)



The voltage of this demonstration Van de Graaff generator is measured

between the charged sphere and ground. Earth's potential is taken to be zero as a reference. The potential of the charged conducting sphere is the same as that of an equal point charge at its center.

Strategy

The potential on the surface will be the same as that of a point charge at the center of the sphere, 12.5 cm away. (The radius of the sphere is 12.5 cm.) We can thus determine the excess charge using the equation

Equation:

$$V = \frac{\text{kQ}}{r}.$$

Solution

Solving for Q and entering known values gives

Equation:

$$\begin{aligned} Q &= \frac{\text{rV}}{k} \\ &= \frac{(0.125\text{ m})\left(100\times10^{3}\text{ V}\right)}{8.99\times10^{9}\text{ N}\cdot\text{m}^2/\text{C}^2} \\ &= 1.39\times10^{-6}\text{ C} = 1.39\ \mu\text{C}. \end{aligned}$$

Discussion

This is a relatively small charge, but it produces a rather large voltage. We have another indication here that it is difficult to store isolated charges.

The voltages in both of these examples could be measured with a meter that compares the measured potential with ground potential. Ground potential is often taken to be zero (instead of taking the potential at infinity to be zero). It is the potential difference between two points that is of importance, and very often there is a tacit assumption that some reference point, such as Earth or a very distant point, is at zero potential. As noted in Electric Potential Energy: Potential Difference, this is analogous to taking sea level as $h = 0$ when considering gravitational potential energy, $\mathrm{PE_g} = \mathrm{mgh}$.

Section Summary

- Electric potential of a point charge is $V = \mathrm{kQ}/r$.
- Electric potential is a scalar, and electric field is a vector. Addition of voltages as numbers gives the voltage due to a combination of point charges, whereas addition of individual fields as vectors gives the total electric field.

Conceptual Questions

Exercise:

Problem:

In what region of space is the potential due to a uniformly charged sphere the same as that of a point charge? In what region does it differ from that of a point charge?

Exercise:

Problem:

Can the potential of a non-uniformly charged sphere be the same as that of a point charge? Explain.

Problems & Exercises

Exercise:

Problem:

A 0.500 cm diameter plastic sphere, used in a static electricity demonstration, has a uniformly distributed 40.0 pC charge on its surface. What is the potential near its surface?

Solution:

144 V

Exercise:

Problem:

What is the potential 0.530×10^{-10} m from a proton (the average distance between the proton and electron in a hydrogen atom)?

Exercise:

Problem:

(a) A sphere has a surface uniformly charged with 1.00 C. At what distance from its center is the potential 5.00 MV? (b) What does your answer imply about the practical aspect of isolating such a large charge?

Solution:

(a) 1.80 km

(b) A charge of 1 C is a very large amount of charge; a sphere of radius 1.80 km is not practical.

Exercise:

Problem:

How far from a $1.00\ \mu\text{C}$ point charge will the potential be 100 V? At what distance will it be 2.00×10^2 V?

Exercise:

Problem:

What are the sign and magnitude of a point charge that produces a potential of -2.00 V at a distance of 1.00 mm?

Solution:

$-2.22 \times 10^{-13} \text{ C}$

Exercise:

Problem:

If the potential due to a point charge is $5.00 \times 10^{2} \text{ V}$ at a distance of 15.0 m, what are the sign and magnitude of the charge?

Exercise:

Problem:

In nuclear fission, a nucleus splits roughly in half. (a) What is the potential $2.00 \times 10^{-14} \text{ m}$ from a fragment that has 46 protons in it? (b) What is the potential energy in MeV of a similarly charged fragment at this distance?

Solution:

(a) $3.31 \times 10^{6} \text{ V}$

(b) 152 MeV

Exercise:

Problem:

A research Van de Graaff generator has a 2.00-m-diameter metal sphere with a charge of 5.00 mC on it. (a) What is the potential near its surface? (b) At what distance from its center is the potential 1.00 MV? (c) An oxygen atom with three missing electrons is released near the Van de Graaff generator. What is its energy in MeV at this distance?

Exercise:

Problem:

An electrostatic paint sprayer has a 0.200-m-diameter metal sphere at a potential of 25.0 kV that repels paint droplets onto a grounded object. (a) What charge is on the sphere? (b) What charge must a 0.100-mg drop of paint have to arrive at the object with a speed of 10.0 m/s?

Solution:

(a) 2.78×10^{-7} C

(b) 2.00×10^{-10} C

Exercise:

Problem:

In one of the classic nuclear physics experiments at the beginning of the 20th century, an alpha particle was accelerated toward a gold nucleus, and its path was substantially deflected by the Coulomb interaction. If the energy of the doubly charged alpha nucleus was 5.00 MeV, how close to the gold nucleus (79 protons) could it come before being deflected?

Exercise:

Problem:

(a) What is the potential between two points situated 10 cm and 20 cm from a $3.0\ \mu C$ point charge? (b) To what location should the point at 20 cm be moved to increase this potential difference by a factor of two?

Exercise:

Problem: Unreasonable Results

(a) What is the final speed of an electron accelerated from rest through a voltage of 25.0 MV by a negatively charged Van de Graaff terminal?

(b) What is unreasonable about this result?

(c) Which assumptions are responsible?

Solution:

(a) $2.96 \times 10^9 \text{ m/s}$

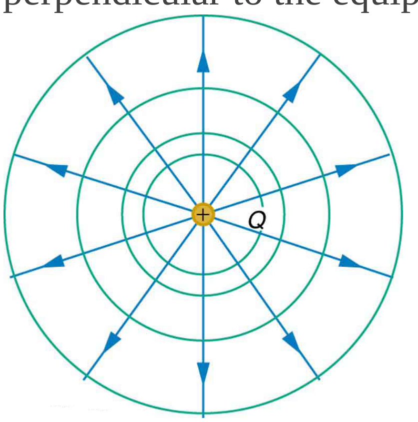
(b) This velocity is far too great. It is faster than the speed of light.

(c) The assumption that the speed of the electron is far less than that of light and that the problem does not require a relativistic treatment produces an answer greater than the speed of light.

Equipotential Lines

- Explain equipotential lines and equipotential surfaces.
- Describe the action of grounding an electrical appliance.
- Compare electric field and equipotential lines.

We can represent electric potentials (voltages) pictorially, just as we drew pictures to illustrate electric fields. Of course, the two are related. Consider [link], which shows an isolated positive point charge and its electric field lines. Electric field lines radiate out from a positive charge and terminate on negative charges. While we use blue arrows to represent the magnitude and direction of the electric field, we use green lines to represent places where the electric potential is constant. These are called **equipotential lines** in two dimensions, or *equipotential surfaces* in three dimensions. The term *equipotential* is also used as a noun, referring to an equipotential line or surface. The potential for a point charge is the same anywhere on an imaginary sphere of radius r surrounding the charge. This is true since the potential for a point charge is given by $V = \mathrm{k}Q/r$ and, thus, has the same value at any point that is a given distance r from the charge. An equipotential sphere is a circle in the two-dimensional view of [link]. Since the electric field lines point radially away from the charge, they are perpendicular to the equipotential lines.



An isolated point charge Q with its electric field lines in blue and equipotential lines

in green. The potential is the same along each equipotential line, meaning that no work is required to move a charge anywhere along one of those lines. Work is needed to move a charge from one equipotential line to another. Equipotential lines are perpendicular to electric field lines in every case.

It is important to note that *equipotential lines are always perpendicular to electric field lines*. No work is required to move a charge along an equipotential, since $\Delta V = 0$. Thus the work is

Equation:

$$W = -\Delta \text{ PE} = -q\Delta V = 0.$$

Work is zero if force is perpendicular to motion. Force is in the same direction as $\mathbf{E}$, so that motion along an equipotential must be perpendicular to $\mathbf{E}$. More precisely, work is related to the electric field by

Equation:

$$W = Fd\cos\theta = qEd\cos\theta = 0.$$

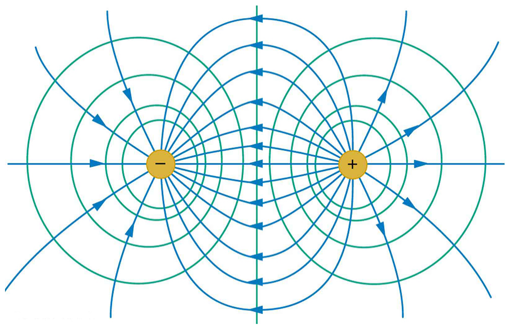
Note that in the above equation, E and F symbolize the magnitudes of the electric field strength and force, respectively. Neither q nor $\mathbf{E}$ nor d is zero, and so $\cos\theta$ must be 0, meaning θ must be 90^o. In other words, motion along an equipotential is perpendicular to $\mathbf{E}$.

One of the rules for static electric fields and conductors is that the electric field must be perpendicular to the surface of any conductor. This implies that a *conductor is an equipotential surface in static situations*. There can be no voltage difference across the surface of a conductor, or charges will flow. One of the uses of this fact is that a conductor can be fixed at zero volts by connecting it to the earth with a good conductor—a process called **grounding**. Grounding can be a useful safety tool. For example, grounding the metal case of an electrical appliance ensures that it is at zero volts relative to the earth.

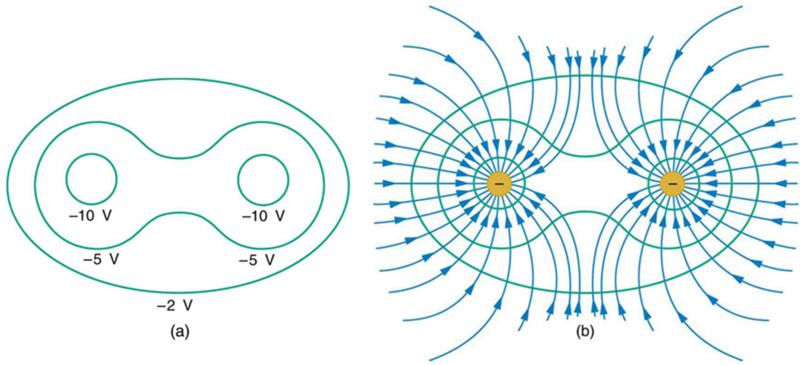
Note:
Grounding
A conductor can be fixed at zero volts by connecting it to the earth with a good conductor—a process called grounding.

Because a conductor is an equipotential, it can replace any equipotential surface. For example, in [link] a charged spherical conductor can replace the point charge, and the electric field and potential surfaces outside of it will be unchanged, confirming the contention that a spherical charge distribution is equivalent to a point charge at its center.

[link] shows the electric field and equipotential lines for two equal and opposite charges. Given the electric field lines, the equipotential lines can be drawn simply by making them perpendicular to the electric field lines. Conversely, given the equipotential lines, as in [link](a), the electric field lines can be drawn by making them perpendicular to the equipotentials, as in [link](b).



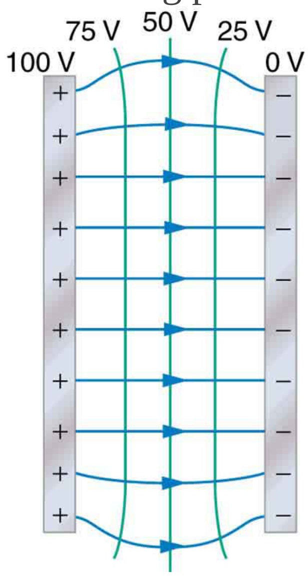
The electric field lines and equipotential lines for two equal but opposite charges. The equipotential lines can be drawn by making them perpendicular to the electric field lines, if those are known. Note that the potential is greatest (most positive) near the positive charge and least (most negative) near the negative charge.



(a) These equipotential lines might be measured with a voltmeter in a laboratory experiment. (b) The corresponding electric field lines are found by drawing them

perpendicular to the equipotentials. Note that these fields are consistent with two equal negative charges.

One of the most important cases is that of the familiar parallel conducting plates shown in [link]. Between the plates, the equipotentials are evenly spaced and parallel. The same field could be maintained by placing conducting plates at the equipotential lines at the potentials shown.



The electric field and equipotential lines between two metal plates.

An important application of electric fields and equipotential lines involves the heart. The heart relies on electrical signals to maintain its rhythm. The movement of electrical signals causes the chambers of the heart to contract and relax. When a person has a heart attack, the movement of these electrical signals may be disturbed. An artificial pacemaker and a

defibrillator can be used to initiate the rhythm of electrical signals. The equipotential lines around the heart, the thoracic region, and the axis of the heart are useful ways of monitoring the structure and functions of the heart. An electrocardiogram (ECG) measures the small electric signals being generated during the activity of the heart. More about the relationship between electric fields and the heart is discussed in Energy Stored in Capacitors.

Note:
PhET Explorations: Charges and Fields
Move point charges around on the playing field and then view the electric field, voltages, equipotential lines, and more. It's colorful, it's dynamic, it's free.
https://phet.colorado.edu/sims/html/charges-and-fields/latest/charges-and-fields_en.html

Section Summary

- An equipotential line is a line along which the electric potential is constant.
- An equipotential surface is a three-dimensional version of equipotential lines.
- Equipotential lines are always perpendicular to electric field lines.
- The process by which a conductor can be fixed at zero volts by connecting it to the earth with a good conductor is called grounding.

Conceptual Questions

Exercise:

Problem:

What is an equipotential line? What is an equipotential surface?

Exercise:

Problem:

Explain in your own words why equipotential lines and surfaces must be perpendicular to electric field lines.

Exercise:

Problem: Can different equipotential lines cross? Explain.

Problems & Exercises

Exercise:

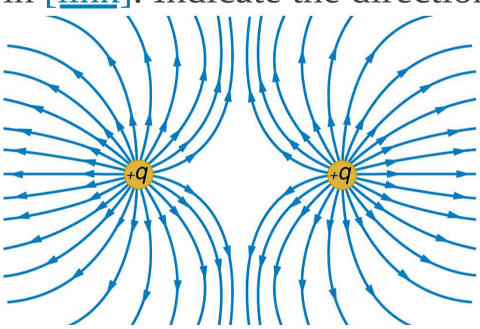
Problem:

(a) Sketch the equipotential lines near a point charge $+q$. Indicate the direction of increasing potential. (b) Do the same for a point charge $-3q$.

Exercise:

Problem:

Sketch the equipotential lines for the two equal positive charges shown in [link]. Indicate the direction of increasing potential.



The electric field near two equal positive charges is directed away from each of the charges.

Exercise:

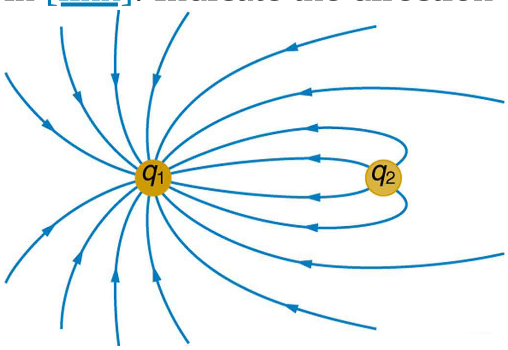
Problem:

[link] shows the electric field lines near two charges q_1 and q_2, the first having a magnitude four times that of the second. Sketch the equipotential lines for these two charges, and indicate the direction of increasing potential.

Exercise:

Problem:

Sketch the equipotential lines a long distance from the charges shown in [link]. Indicate the direction of increasing potential.



The electric field near two charges.

Exercise:

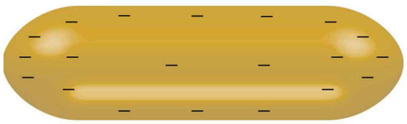
Problem:

Sketch the equipotential lines in the vicinity of two opposite charges, where the negative charge is three times as great in magnitude as the positive. See [link] for a similar situation. Indicate the direction of increasing potential.

Exercise:

Problem:

Sketch the equipotential lines in the vicinity of the negatively charged conductor in [link]. How will these equipotentials look a long distance from the object?

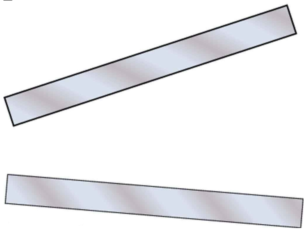


A negatively charged conductor.

Exercise:

Problem:

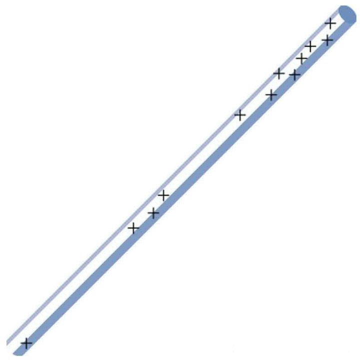
Sketch the equipotential lines surrounding the two conducting plates shown in [link], given the top plate is positive and the bottom plate has an equal amount of negative charge. Be certain to indicate the distribution of charge on the plates. Is the field strongest where the plates are closest? Why should it be?



Exercise:

Problem:

(a) Sketch the electric field lines in the vicinity of the charged insulator in [link]. Note its non-uniform charge distribution. (b) Sketch equipotential lines surrounding the insulator. Indicate the direction of increasing potential.



A charged insulating rod such as might be used in a classroom demonstration.

Exercise:

Problem:

The naturally occurring charge on the ground on a fine day out in the open country is $-1.00\ \mathrm{nC/m^2}$. (a) What is the electric field relative to ground at a height of 3.00 m? (b) Calculate the electric potential at this height. (c) Sketch electric field and equipotential lines for this scenario.

Exercise:

Problem:

The lesser electric ray (*Narcine bancroftii*) maintains an incredible charge on its head and a charge equal in magnitude but opposite in sign on its tail ([link]). (a) Sketch the equipotential lines surrounding the ray. (b) Sketch the equipotentials when the ray is near a ship with a conducting surface. (c) How could this charge distribution be of use to the ray?

Lesser electric ray (*Narcine bancroftii*) (credit: National Oceanic and Atmospheric Administration, NOAA's Fisheries Collection).

Glossary

equipotential line
: a line along which the electric potential is constant

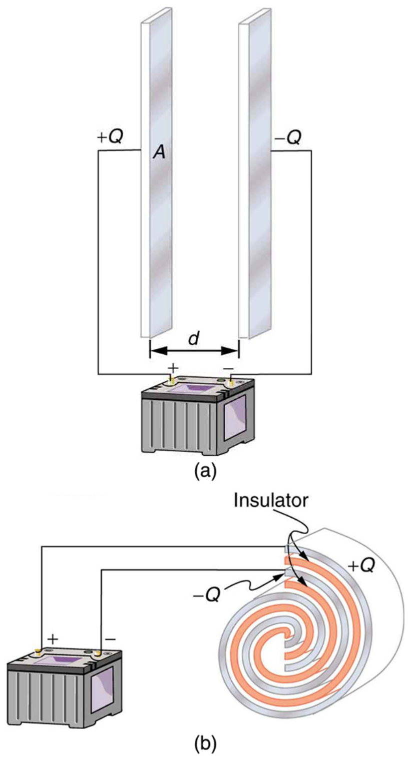
grounding
: fixing a conductor at zero volts by connecting it to the earth or ground

Capacitors and Dielectrics

- Describe the action of a capacitor and define capacitance.
- Explain parallel plate capacitors and their capacitances.
- Discuss the process of increasing the capacitance of a dielectric.
- Determine capacitance given charge and voltage.

A **capacitor** is a device used to store electric charge. Capacitors have applications ranging from filtering static out of radio reception to energy storage in heart defibrillators. Typically, commercial capacitors have two conducting parts close to one another, but not touching, such as those in [link]. (Most of the time an insulator is used between the two plates to provide separation—see the discussion on dielectrics below.) When battery terminals are connected to an initially uncharged capacitor, equal amounts of positive and negative charge, $+Q$ and $-Q$, are separated into its two plates. The capacitor remains neutral overall, but we refer to it as storing a charge Q in this circumstance.

Note:
Capacitor
A capacitor is a device used to store electric charge.

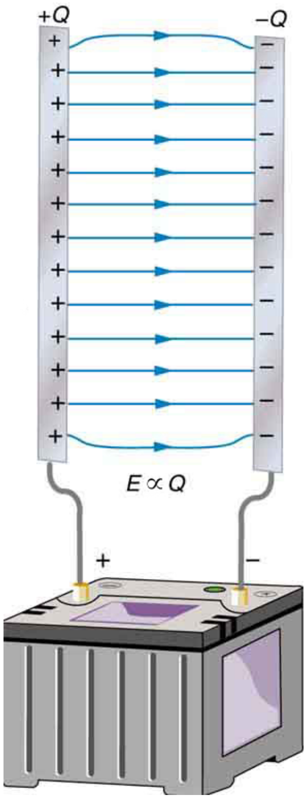


Both capacitors shown here were initially uncharged before being connected to a battery. They now have separated charges of $+Q$ and $-Q$ on their two halves. (a) A parallel plate capacitor. (b) A rolled capacitor with an insulating material between its two conducting sheets.

The amount of charge Q a *capacitor* can store depends on two major factors—the voltage applied and the capacitor’s physical characteristics, such as its size.

Note:
The Amount of Charge Q a Capacitor Can Store
The amount of charge Q a *capacitor* can store depends on two major factors—the voltage applied and the capacitor’s physical characteristics, such as its size.

A system composed of two identical, parallel conducting plates separated by a distance, as in [link], is called a **parallel plate capacitor**. It is easy to see the relationship between the voltage and the stored charge for a parallel plate capacitor, as shown in [link]. Each electric field line starts on an individual positive charge and ends on a negative one, so that there will be more field lines if there is more charge. (Drawing a single field line per charge is a convenience, only. We can draw many field lines for each charge, but the total number is proportional to the number of charges.) The electric field strength is, thus, directly proportional to Q.



Electric field lines in this parallel plate capacitor, as always, start on positive charges and end on negative charges. Since the electric field strength is proportional to the density of field lines, it is also proportional to the amount

of charge on
the capacitor.

The field is proportional to the charge:

Equation:

$$E\propto Q,$$

where the symbol $\propto$ means “proportional to.” From the discussion in Electric Potential in a Uniform Electric Field, we know that the voltage across parallel plates is $V=\text{Ed}$. Thus,

Equation:

$$V\propto E.$$

It follows, then, that $V\propto Q$, and conversely,

Equation:

$$Q\propto V.$$

This is true in general: The greater the voltage applied to any capacitor, the greater the charge stored in it.

Different capacitors will store different amounts of charge for the same applied voltage, depending on their physical characteristics. We define their **capacitance** C to be such that the charge Q stored in a capacitor is proportional to C. The charge stored in a capacitor is given by

Equation:

$$Q=\text{CV}.$$

This equation expresses the two major factors affecting the amount of charge stored. Those factors are the physical characteristics of the capacitor,

C, and the voltage, V. Rearranging the equation, we see that *capacitance* C *is the amount of charge stored per volt*, or
Equation:

$$C = \frac{Q}{V}.$$

Note:
Capacitance
Capacitance C is the amount of charge stored per volt, or
Equation:

$$C = \frac{Q}{V}.$$

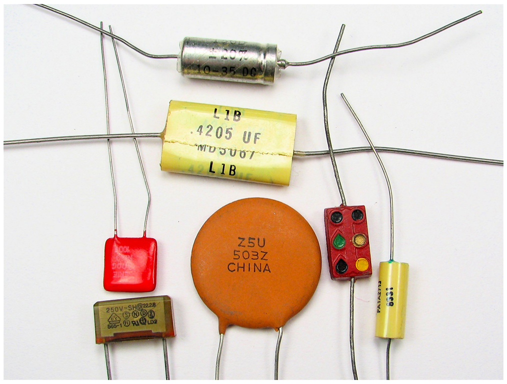
The unit of capacitance is the farad (F), named for Michael Faraday (1791–1867), an English scientist who contributed to the fields of electromagnetism and electrochemistry. Since capacitance is charge per unit voltage, we see that a farad is a coulomb per volt, or
Equation:

$$1\text{ F} = \frac{1\text{ C}}{1\text{ V}}.$$

A 1-farad capacitor would be able to store 1 coulomb (a very large amount of charge) with the application of only 1 volt. One farad is, thus, a very large capacitance. Typical capacitors range from fractions of a picofarad $\left(1\text{ pF} = 10^{-12}\text{ F}\right)$ to millifarads $\left(1\text{ mF} = 10^{-3}\text{ F}\right)$.

[link] shows some common capacitors. Capacitors are primarily made of ceramic, glass, or plastic, depending upon purpose and size. Insulating

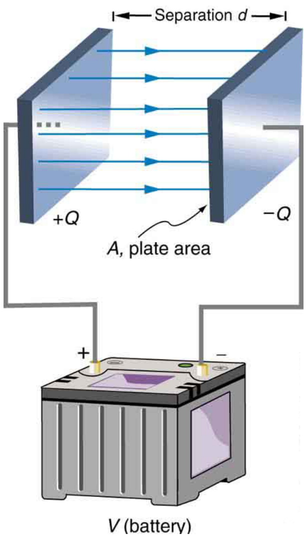
materials, called dielectrics, are commonly used in their construction, as discussed below.



Some typical capacitors. Size and value of capacitance are not necessarily related. (credit: Windell Oskay)

Parallel Plate Capacitor

The parallel plate capacitor shown in [link] has two identical conducting plates, each having a surface area A, separated by a distance d (with no material between the plates). When a voltage V is applied to the capacitor, it stores a charge Q, as shown. We can see how its capacitance depends on A and d by considering the characteristics of the Coulomb force. We know that like charges repel, unlike charges attract, and the force between charges decreases with distance. So it seems quite reasonable that the bigger the plates are, the more charge they can store—because the charges can spread out more. Thus C should be greater for larger A. Similarly, the closer the plates are together, the greater the attraction of the opposite charges on them. So C should be greater for smaller d.



Parallel plate capacitor with plates separated by a distance d. Each plate has an area A.

It can be shown that for a parallel plate capacitor there are only two factors (A and d) that affect its capacitance C. The capacitance of a parallel plate capacitor in equation form is given by

Equation:

$$C = \varepsilon_0 \frac{A}{d}.$$

Note:
Capacitance of a Parallel Plate Capacitor
Equation:

$$C=\varepsilon_0\frac{A}{d}$$

A is the area of one plate in square meters, and d is the distance between the plates in meters. The constant ε_0 is the permittivity of free space; its numerical value in SI units is $\varepsilon_0 = 8.85 \times 10^{-12}\ \mathrm{F/m}$. The units of F/m are equivalent to $\mathrm{C^2/N \cdot m^2}$. The small numerical value of ε_0 is related to the large size of the farad. A parallel plate capacitor must have a large area to have a capacitance approaching a farad. (Note that the above equation is valid when the parallel plates are separated by air or free space. When another material is placed between the plates, the equation is modified, as discussed below.)

Example:
Capacitance and Charge Stored in a Parallel Plate Capacitor
(a) What is the capacitance of a parallel plate capacitor with metal plates, each of area $1.00\ \mathrm{m}^2$, separated by 1.00 mm? (b) What charge is stored in this capacitor if a voltage of 3.00×10^3 V is applied to it?
Strategy
Finding the capacitance C is a straightforward application of the equation $C=\varepsilon_0 A/d$. Once C is found, the charge stored can be found using the equation $Q = \mathrm{CV}$.
Solution for (a)
Entering the given values into the equation for the capacitance of a parallel plate capacitor yields
Equation:

$$\begin{aligned} C &= \varepsilon_0\frac{A}{d} = 8.85 \times 10^{-12}\,\frac{\mathrm{F}}{\mathrm{m}}\ \frac{1.00\ \mathrm{m}^2}{1.00\times10^{-3}\ \mathrm{m}} \\ &= 8.85 \times 10^{-9}\ \mathrm{F} = 8.85\ \mathrm{nF}. \end{aligned}$$

Discussion for (a)

This small value for the capacitance indicates how difficult it is to make a device with a large capacitance. Special techniques help, such as using very large area thin foils placed close together.

Solution for (b)

The charge stored in any capacitor is given by the equation $Q = \text{CV}$. Entering the known values into this equation gives

Equation:

$$\begin{aligned} Q &= CV = \left(8.85 \times 10^{-9}\ \text{F}\right)\left(3.00 \times 10^{3}\ \text{V}\right) \\ &= 26.6\ \mu\text{C}. \end{aligned}$$

Discussion for (b)

This charge is only slightly greater than those found in typical static electricity. Since air breaks down at about $3.00 \times 10^{6}\ \text{V/m}$, more charge cannot be stored on this capacitor by increasing the voltage.

Another interesting biological example dealing with electric potential is found in the cell's plasma membrane. The membrane sets a cell off from its surroundings and also allows ions to selectively pass in and out of the cell. There is a potential difference across the membrane of about $–70\ \text{mV}$. This is due to the mainly negatively charged ions in the cell and the predominance of positively charged sodium (Na^{+}) ions outside. Things change when a nerve cell is stimulated. Na^{+} ions are allowed to pass through the membrane into the cell, producing a positive membrane potential—the nerve signal. The cell membrane is about 7 to 10 nm thick. An approximate value of the electric field across it is given by

Equation:

$$E = \frac{V}{d} = \frac{–70 \times 10^{-3}\ \text{V}}{8 \times 10^{-9}\ \text{m}} = –9 \times 10^{6}\ \text{V/m}.$$

This electric field is enough to cause a breakdown in air.

Dielectric

The previous example highlights the difficulty of storing a large amount of charge in capacitors. If d is made smaller to produce a larger capacitance, then the maximum voltage must be reduced proportionally to avoid breakdown (since $E = V/d$). An important solution to this difficulty is to put an insulating material, called a **dielectric**, between the plates of a capacitor and allow d to be as small as possible. Not only does the smaller d make the capacitance greater, but many insulators can withstand greater electric fields than air before breaking down.

There is another benefit to using a dielectric in a capacitor. Depending on the material used, the capacitance is greater than that given by the equation $C = \varepsilon_0 \frac{A}{d}$ by a factor κ, called the *dielectric constant*. A parallel plate capacitor with a dielectric between its plates has a capacitance given by

Equation:

$$C = \kappa \varepsilon_0 \frac{A}{d} \text{ (parallel plate capacitor with dielectric).}$$

Values of the dielectric constant κ for various materials are given in [link]. Note that κ for vacuum is exactly 1, and so the above equation is valid in that case, too. If a dielectric is used, perhaps by placing Teflon between the plates of the capacitor in [link], then the capacitance is greater by the factor κ, which for Teflon is 2.1.

Note:

Take-Home Experiment: Building a Capacitor

How large a capacitor can you make using a chewing gum wrapper? The plates will be the aluminum foil, and the separation (dielectric) in between will be the paper.

Material	Dielectric constant κ	Dielectric strength (V/m)
Vacuum	1.00000	—
Air	1.00059	3×10^6
Bakelite	4.9	24×10^6
Fused quartz	3.78	8×10^6
Neoprene rubber	6.7	12×10^6
Nylon	3.4	14×10^6
Paper	3.7	16×10^6
Polystyrene	2.56	24×10^6
Pyrex glass	5.6	14×10^6
Silicon oil	2.5	15×10^6
Strontium titanate	233	8×10^6
Teflon	2.1	60×10^6
Water	80	—

Dielectric Constants and Dielectric Strengths for Various Materials at 20ºC

Note also that the dielectric constant for air is very close to 1, so that air-filled capacitors act much like those with vacuum between their plates *except* that the air can become conductive if the electric field strength

becomes too great. (Recall that $E = V/d$ for a parallel plate capacitor.) Also shown in [link] are maximum electric field strengths in V/m, called **dielectric strengths**, for several materials. These are the fields above which the material begins to break down and conduct. The dielectric strength imposes a limit on the voltage that can be applied for a given plate separation. For instance, in [link], the separation is 1.00 mm, and so the voltage limit for air is

Equation:

$$\begin{aligned} V &= E \cdot d \\ &= (3 \times 10^6 \text{ V/m})(1.00 \times 10^{-3} \text{ m}) \\ &= 3000 \text{ V.} \end{aligned}$$

However, the limit for a 1.00 mm separation filled with Teflon is 60,000 V, since the dielectric strength of Teflon is 60×10^6 V/m. So the same capacitor filled with Teflon has a greater capacitance and can be subjected to a much greater voltage. Using the capacitance we calculated in the above example for the air-filled parallel plate capacitor, we find that the Teflon-filled capacitor can store a maximum charge of

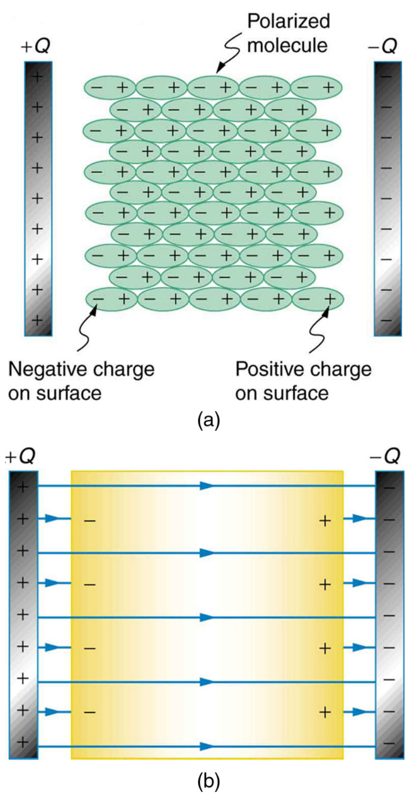
Equation:

$$\begin{aligned} Q &= CV \\ &= \kappa C_{\text{air}} V \\ &= (2.1)(8.85 \text{ nF})(6.0 \times 10^4 \text{ V}) \\ &= 1.1 \text{ mC.} \end{aligned}$$

This is 42 times the charge of the same air-filled capacitor.

Note:
Dielectric Strength
The maximum electric field strength above which an insulating material begins to break down and conduct is called its dielectric strength.

Microscopically, how does a dielectric increase capacitance? Polarization of the insulator is responsible. The more easily it is polarized, the greater its dielectric constant κ. Water, for example, is a **polar molecule** because one end of the molecule has a slight positive charge and the other end has a slight negative charge. The polarity of water causes it to have a relatively large dielectric constant of 80. The effect of polarization can be best explained in terms of the characteristics of the Coulomb force. [link] shows the separation of charge schematically in the molecules of a dielectric material placed between the charged plates of a capacitor. The Coulomb force between the closest ends of the molecules and the charge on the plates is attractive and very strong, since they are very close together. This attracts more charge onto the plates than if the space were empty and the opposite charges were a distance d away.



(a) The molecules in the insulating material between

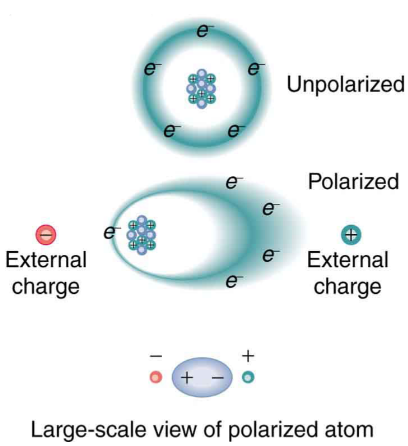
the plates of a capacitor are polarized by the charged plates. This produces a layer of opposite charge on the surface of the dielectric that attracts more charge onto the plate, increasing its capacitance. (b) The dielectric reduces the electric field strength inside the capacitor, resulting in a smaller voltage between the plates for the same charge. The capacitor stores the same charge for a smaller voltage, implying that it has a larger capacitance because of the dielectric.

Another way to understand how a dielectric increases capacitance is to consider its effect on the electric field inside the capacitor. [link](b) shows the electric field lines with a dielectric in place. Since the field lines end on charges in the dielectric, there are fewer of them going from one side of the capacitor to the other. So the electric field strength is less than if there were

a vacuum between the plates, even though the same charge is on the plates. The voltage between the plates is $V = \mathrm{Ed}$, so it too is reduced by the dielectric. Thus there is a smaller voltage V for the same charge Q; since $C = Q/V$, the capacitance C is greater.

The dielectric constant is generally defined to be $\kappa = E_0/E$, or the ratio of the electric field in a vacuum to that in the dielectric material, and is intimately related to the polarizability of the material.

Note:
Things Great and Small
The Submicroscopic Origin of Polarization
Polarization is a separation of charge within an atom or molecule. As has been noted, the planetary model of the atom pictures it as having a positive nucleus orbited by negative electrons, analogous to the planets orbiting the Sun. Although this model is not completely accurate, it is very helpful in explaining a vast range of phenomena and will be refined elsewhere, such as in Atomic Physics. The submicroscopic origin of polarization can be modeled as shown in [link].



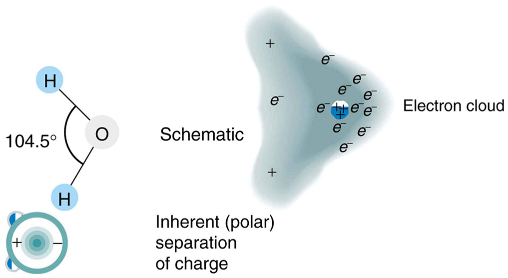
Artist’s conception of a polarized atom. The orbits of

electrons around the nucleus are shifted slightly by the external charges (shown exaggerated). The resulting separation of charge within the atom means that it is polarized. Note that the unlike charge is now closer to the external charges, causing the polarization.

We will find in Atomic Physics that the orbits of electrons are more properly viewed as electron clouds with the density of the cloud related to the probability of finding an electron in that location (as opposed to the definite locations and paths of planets in their orbits around the Sun). This cloud is shifted by the Coulomb force so that the atom on average has a separation of charge. Although the atom remains neutral, it can now be the source of a Coulomb force, since a charge brought near the atom will be closer to one type of charge than the other.

Some molecules, such as those of water, have an inherent separation of charge and are thus called polar molecules. [link] illustrates the separation of charge in a water molecule, which has two hydrogen atoms and one oxygen atom (H_2O). The water molecule is not symmetric—the hydrogen atoms are repelled to one side, giving the molecule a boomerang shape. The electrons in a water molecule are more concentrated around the more highly charged oxygen nucleus than around the hydrogen nuclei. This makes the oxygen end of the molecule slightly negative and leaves the hydrogen ends slightly positive. The inherent separation of charge in polar molecules

makes it easier to align them with external fields and charges. Polar molecules therefore exhibit greater polarization effects and have greater dielectric constants. Those who study chemistry will find that the polar nature of water has many effects. For example, water molecules gather ions much more effectively because they have an electric field and a separation of charge to attract charges of both signs. Also, as brought out in the previous chapter, polar water provides a shield or screening of the electric fields in the highly charged molecules of interest in biological systems.



Artist’s conception of a water molecule. There is an inherent separation of charge, and so water is a polar molecule. Electrons in the molecule are attracted to the oxygen nucleus and leave an excess of positive charge near the two hydrogen nuclei. (Note that the schematic on the right is a rough illustration of the distribution of electrons in the water molecule. It does not show the actual numbers of protons and electrons involved in the structure.)

Note:
PhET Explorations: Capacitor Lab
Explore how a capacitor works! Change the size of the plates and add a dielectric to see the effect on capacitance. Change the voltage and see charges built up on the plates. Observe the electric field in the capacitor. Measure the voltage and the electric field.

Capacitor Lab

Section Summary

- A capacitor is a device used to store charge.
- The amount of charge Q a capacitor can store depends on two major factors—the voltage applied and the capacitor's physical characteristics, such as its size.
- The capacitance C is the amount of charge stored per volt, or
  Equation:
  $$C = \frac{Q}{V}.$$
- The capacitance of a parallel plate capacitor is $C = \varepsilon_0 \frac{A}{d}$, when the plates are separated by air or free space. ε_0 is called the permittivity of free space.
- A parallel plate capacitor with a dielectric between its plates has a capacitance given by
  Equation:

$$C = \kappa\varepsilon_0 \frac{A}{d},$$

where κ is the dielectric constant of the material.
- The maximum electric field strength above which an insulating material begins to break down and conduct is called dielectric strength.

Conceptual Questions

Exercise:

Problem:

Does the capacitance of a device depend on the applied voltage? What about the charge stored in it?

Exercise:

Problem:

Use the characteristics of the Coulomb force to explain why capacitance should be proportional to the plate area of a capacitor. Similarly, explain why capacitance should be inversely proportional to the separation between plates.

Exercise:

Problem:

Give the reason why a dielectric material increases capacitance compared with what it would be with air between the plates of a capacitor. What is the independent reason that a dielectric material also allows a greater voltage to be applied to a capacitor? (The dielectric thus increases C and permits a greater V.)

Exercise:

Problem:

How does the polar character of water molecules help to explain water's relatively large dielectric constant? ([link])

Exercise:

Problem:

Sparks will occur between the plates of an air-filled capacitor at lower voltage when the air is humid than when dry. Explain why, considering the polar character of water molecules.

Exercise:

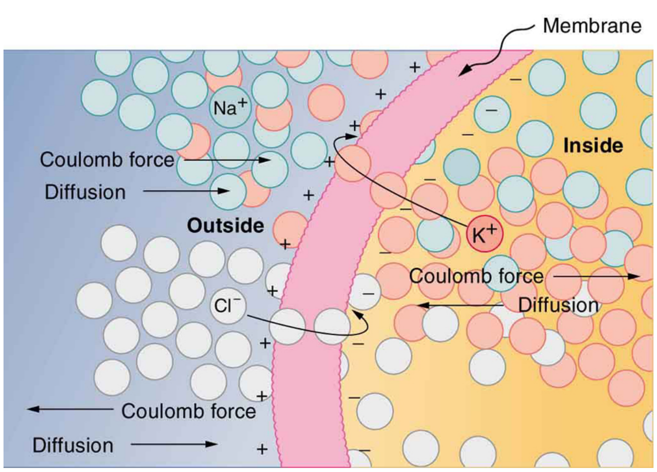
Problem:

Water has a large dielectric constant, but it is rarely used in capacitors. Explain why.

Exercise:

Problem:

Membranes in living cells, including those in humans, are characterized by a separation of charge across the membrane. Effectively, the membranes are thus charged capacitors with important functions related to the potential difference across the membrane. Is energy required to separate these charges in living membranes and, if so, is its source the metabolization of food energy or some other source?



The semipermeable membrane of a cell has different concentrations of ions inside and out. Diffusion moves the K^+ (potassium) and Cl^- (chloride) ions in the directions shown, until the Coulomb force halts further transfer. This results in a layer of positive charge on the outside, a layer of negative charge on the inside, and thus a voltage across the cell membrane. The membrane is normally impermeable to Na^+ (sodium ions).

Problems & Exercises

Exercise:

Problem:

What charge is stored in a $180\ \mu F$ capacitor when 120 V is applied to it?

Solution:

$21.6\ mC$

Exercise:

Problem:

Find the charge stored when 5.50 V is applied to an 8.00 pF capacitor.

Exercise:

Problem: What charge is stored in the capacitor in [link]?

Solution:

$80.0\ \mathrm{mC}$

Exercise:

Problem:

Calculate the voltage applied to a $2.00\ \mu\mathrm{F}$ capacitor when it holds $3.10\ \mu\mathrm{C}$ of charge.

Exercise:

Problem:

What voltage must be applied to an 8.00 nF capacitor to store 0.160 mC of charge?

Solution:

20.0 kV

Exercise:

Problem:

What capacitance is needed to store $3.00\ \mu\mathrm{C}$ of charge at a voltage of 120 V?

Exercise:

Problem:

What is the capacitance of a large Van de Graaff generator's terminal, given that it stores 8.00 mC of charge at a voltage of 12.0 MV?

Solution:

$667\ \mathrm{pF}$

Exercise:

Problem:

Find the capacitance of a parallel plate capacitor having plates of area $5.00\ \text{m}^2$ that are separated by 0.100 mm of Teflon.

Exercise:

Problem:

(a)What is the capacitance of a parallel plate capacitor having plates of area $1.50\ \text{m}^2$ that are separated by 0.0200 mm of neoprene rubber? (b) What charge does it hold when 9.00 V is applied to it?

Solution:

(a) $4.4\ \mu\text{F}$

(b) $4.0 \times 10^{-5}\ \text{C}$

Exercise:

Problem: Integrated Concepts

A prankster applies 450 V to an $80.0\ \mu\text{F}$ capacitor and then tosses it to an unsuspecting victim. The victim's finger is burned by the discharge of the capacitor through 0.200 g of flesh. What is the temperature increase of the flesh? Is it reasonable to assume no phase change?

Exercise:

Problem:Unreasonable Results

(a) A certain parallel plate capacitor has plates of area $4.00\ \text{m}^2$, separated by 0.0100 mm of nylon, and stores 0.170 C of charge. What is the applied voltage? (b) What is unreasonable about this result? (c) Which assumptions are responsible or inconsistent?

Solution:

(a) 14.2 kV

(b) The voltage is unreasonably large, more than 100 times the breakdown voltage of nylon.

(c) The assumed charge is unreasonably large and cannot be stored in a capacitor of these dimensions.

Glossary

capacitor
: a device that stores electric charge

capacitance
: amount of charge stored per unit volt

dielectric
: an insulating material

dielectric strength
: the maximum electric field above which an insulating material begins to break down and conduct

parallel plate capacitor
: two identical conducting plates separated by a distance

polar molecule
: a molecule with inherent separation of charge

Capacitors in Series and Parallel

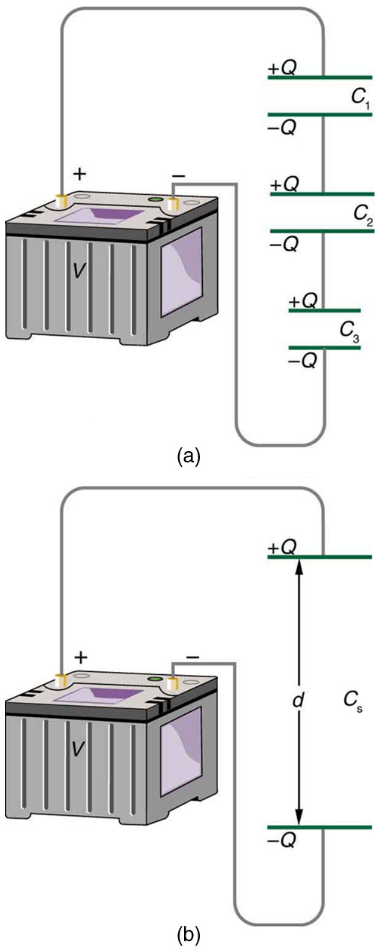
- Derive expressions for total capacitance in series and in parallel.
- Identify series and parallel parts in the combination of connection of capacitors.
- Calculate the effective capacitance in series and parallel given individual capacitances.

Several capacitors may be connected together in a variety of applications. Multiple connections of capacitors act like a single equivalent capacitor. The total capacitance of this equivalent single capacitor depends both on the individual capacitors and how they are connected. There are two simple and common types of connections, called *series* and *parallel*, for which we can easily calculate the total capacitance. Certain more complicated connections can also be related to combinations of series and parallel.

Capacitance in Series

[link](a) shows a series connection of three capacitors with a voltage applied. As for any capacitor, the capacitance of the combination is related to charge and voltage by $C=\frac{Q}{V}$.

Note in [link] that opposite charges of magnitude Q flow to either side of the originally uncharged combination of capacitors when the voltage V is applied. Conservation of charge requires that equal-magnitude charges be created on the plates of the individual capacitors, since charge is only being separated in these originally neutral devices. The end result is that the combination resembles a single capacitor with an effective plate separation greater than that of the individual capacitors alone. (See [link](b).) Larger plate separation means smaller capacitance. It is a general feature of series connections of capacitors that the total capacitance is less than any of the individual capacitances.



(a) Capacitors connected in series. The magnitude of the charge on each plate is Q. (b) An equivalent capacitor has a larger plate separation d. Series connections produce a total capacitance that is less than that of

any of the individual capacitors.

We can find an expression for the total capacitance by considering the voltage across the individual capacitors shown in [link]. Solving $C=\frac{Q}{V}$ for V gives $V=\frac{Q}{C}$. The voltages across the individual capacitors are thus $V_1=\frac{Q}{C_1}$, $V_2=\frac{Q}{C_2}$, and $V_3=\frac{Q}{C_3}$. The total voltage is the sum of the individual voltages:

Equation:

$$V=V_1+V_2+V_3.$$

Now, calling the total capacitance C_S for series capacitance, consider that

Equation:

$$V=\frac{Q}{C_\mathrm{S}}=V_1+V_2+V_3.$$

Entering the expressions for V_1, V_2, and V_3, we get

Equation:

$$\frac{Q}{C_\mathrm{S}}=\frac{Q}{C_1}+\frac{Q}{C_2}+\frac{Q}{C_3}.$$

Canceling the Qs, we obtain the equation for the total capacitance in series C_S to be

Equation:

$$\frac{1}{C_\mathrm{S}}=\frac{1}{C_1}+\frac{1}{C_2}+\frac{1}{C_3}+\ldots,$$

where “...” indicates that the expression is valid for any number of capacitors connected in series. An expression of this form always results in a total capacitance C_{S} that is less than any of the individual capacitances C_1, C_2, ..., as the next example illustrates.

Note:

Total Capacitance in Series, C_{s}

Total capacitance in series: $\frac{1}{C_{\mathrm{S}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + ...$

Example:

What Is the Series Capacitance?

Find the total capacitance for three capacitors connected in series, given their individual capacitances are 1.000, 5.000, and 8.000 μF.

Strategy

With the given information, the total capacitance can be found using the equation for capacitance in series.

Solution

Entering the given capacitances into the expression for $\frac{1}{C_{\mathrm{S}}}$ gives $\frac{1}{C_S} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}$.

Equation:

$$\frac{1}{C_{\mathrm{S}}} = \frac{1}{1.000\ \mu\mathrm{F}} + \frac{1}{5.000\ \mu\mathrm{F}} + \frac{1}{8.000\ \mu\mathrm{F}} = \frac{1.325}{\mu\mathrm{F}}$$

Inverting to find C_{S} yields $C_{\mathrm{S}} = \frac{\mu\mathrm{F}}{1.325} = 0.755\ \mu\mathrm{F}$.

Discussion

The total series capacitance C_{s} is less than the smallest individual capacitance, as promised. In series connections of capacitors, the sum is less than the parts. In fact, it is less than any individual. Note that it is sometimes possible, and more convenient, to solve an equation like the above by finding the least common denominator, which in this case (showing only whole-number calculations) is 40. Thus,

Equation:

$$\frac{1}{C_\mathrm{S}} = \frac{40}{40\ \mu\mathrm{F}} + \frac{8}{40\ \mu\mathrm{F}} + \frac{5}{40\ \mu\mathrm{F}} = \frac{53}{40\ \mu\mathrm{F}},$$

so that

Equation:

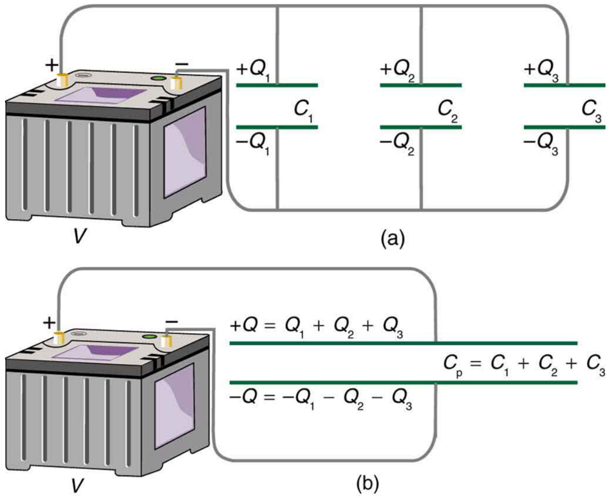
$$C_\mathrm{S} = \frac{40\ \mu\mathrm{F}}{53} = 0.755\ \mu\mathrm{F}.$$

Capacitors in Parallel

[link](a) shows a parallel connection of three capacitors with a voltage applied. Here the total capacitance is easier to find than in the series case. To find the equivalent total capacitance C_p, we first note that the voltage across each capacitor is V, the same as that of the source, since they are connected directly to it through a conductor. (Conductors are equipotentials, and so the voltage across the capacitors is the same as that across the voltage source.) Thus the capacitors have the same charges on them as they would have if connected individually to the voltage source. The total charge Q is the sum of the individual charges:

Equation:

$$Q = Q_1 + Q_2 + Q_3.$$



(a) Capacitors in parallel. Each is connected directly to the voltage source just as if it were all alone, and so the total capacitance in parallel is just the sum of the individual capacitances. (b) The equivalent capacitor has a larger plate area and can therefore hold more charge than the individual capacitors.

Using the relationship $Q = \mathrm{CV}$, we see that the total charge is $Q = C_\mathrm{p}V$, and the individual charges are $Q_1 = C_1V$, $Q_2 = C_2V$, and $Q_3 = C_3V$. Entering these into the previous equation gives

Equation:

$$C_\mathrm{p}V = C_1V + C_2V + C_3V.$$

Canceling V from the equation, we obtain the equation for the total capacitance in parallel C_p:

Equation:

$$C_\text{p} = C_1 + C_2 + C_3 + \text{....}$$

Total capacitance in parallel is simply the sum of the individual capacitances. (Again the "..." indicates the expression is valid for any number of capacitors connected in parallel.) So, for example, if the capacitors in the example above were connected in parallel, their capacitance would be

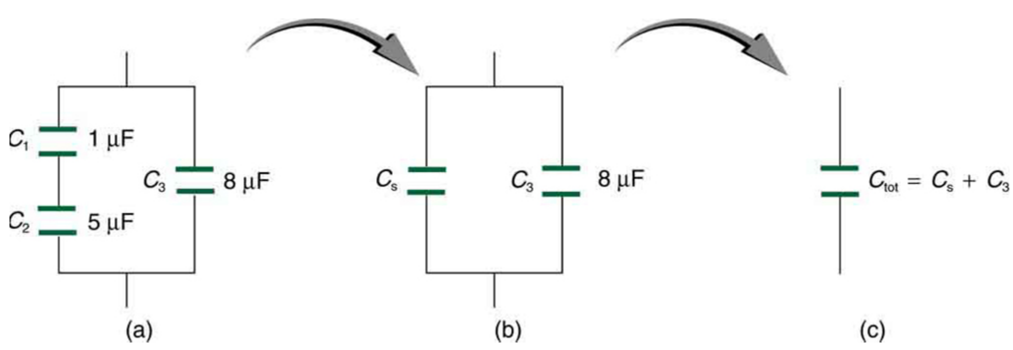
Equation:

$$C_\text{p} = 1.000\ \mu\text{F} + 5.000\ \mu\text{F} + 8.000\ \mu\text{F} = 14.000\ \mu\text{F}.$$

The equivalent capacitor for a parallel connection has an effectively larger plate area and, thus, a larger capacitance, as illustrated in [link](b).

> **Note:**
> Total Capacitance in Parallel, C_p
> Total capacitance in parallel $C_\text{p} = C_1 + C_2 + C_3 + \text{...}$

More complicated connections of capacitors can sometimes be combinations of series and parallel. (See [link].) To find the total capacitance of such combinations, we identify series and parallel parts, compute their capacitances, and then find the total.



(a) This circuit contains both series and parallel connections of capacitors. See [link] for the calculation of the overall capacitance of the circuit. (b) C_1 and C_2 are in series; their equivalent capacitance C_S is less than either of them. (c) Note that C_S is in parallel with C_3. The total capacitance is, thus, the sum of C_S and C_3.

Example:

A Mixture of Series and Parallel Capacitance

Find the total capacitance of the combination of capacitors shown in [link]. Assume the capacitances in [link] are known to three decimal places ($C_1 = 1.000\ \mu F$, $C_2 = 5.000\ \mu F$, and $C_3 = 8.000\ \mu F$), and round your answer to three decimal places.

Strategy

To find the total capacitance, we first identify which capacitors are in series and which are in parallel. Capacitors C_1 and C_2 are in series. Their combination, labeled C_S in the figure, is in parallel with C_3.

Solution

Since C_1 and C_2 are in series, their total capacitance is given by $\frac{1}{C_S} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}$. Entering their values into the equation gives

Equation:

$$\frac{1}{C_S} = \frac{1}{C_1} + \frac{1}{C_2} = \frac{1}{1.000\ \mu F} + \frac{1}{5.000\ \mu F} = \frac{1.200}{\mu F}.$$

Inverting gives

Equation:

$$C_S = 0.833\ \mu F.$$

This equivalent series capacitance is in parallel with the third capacitor; thus, the total is the sum

Equation:

$$\begin{aligned} C_{\text{tot}} &= C_{\text{S}} + C_{\text{S}} \\ &= 0.833\ \mu\text{F} + 8.000\ \mu\text{F} \\ &= 8.833\ \mu\text{F}. \end{aligned}$$

Discussion

This technique of analyzing the combinations of capacitors piece by piece until a total is obtained can be applied to larger combinations of capacitors.

Section Summary

- Total capacitance in series $\frac{1}{C_{\text{S}}} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \ldots$
- Total capacitance in parallel $C_{\text{p}} = C_1 + C_2 + C_3 + \ldots$
- If a circuit contains a combination of capacitors in series and parallel, identify series and parallel parts, compute their capacitances, and then find the total.

Conceptual Questions

Exercise:

Problem:

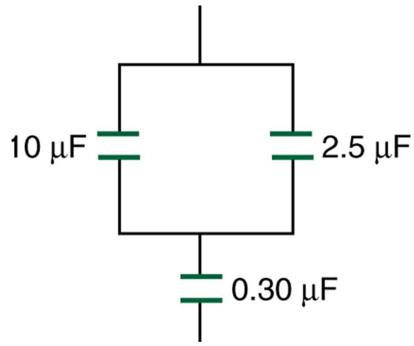
If you wish to store a large amount of energy in a capacitor bank, would you connect capacitors in series or parallel? Explain.

Problems & Exercises

Exercise:

Problem:

Find the total capacitance of the combination of capacitors in [link].



A combination of series and parallel connections of capacitors.

Solution:

0.293 μF

Exercise:

Problem:

Suppose you want a capacitor bank with a total capacitance of 0.750 F and you possess numerous 1.50 mF capacitors. What is the smallest number you could hook together to achieve your goal, and how would you connect them?

Exercise:

Problem:

What total capacitances can you make by connecting a 5.00 μF and an 8.00 μF capacitor together?

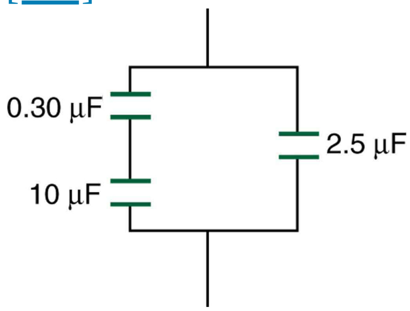
Solution:

3.08 μF in series combination, 13.0 μF in parallel combination

Exercise:

Problem:

Find the total capacitance of the combination of capacitors shown in [link].



A combination of series and parallel connections of capacitors.

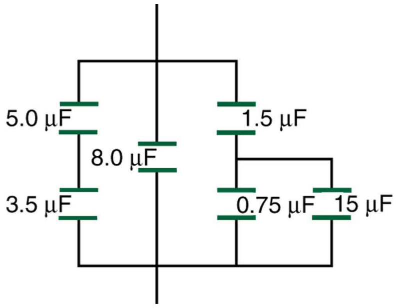
Solution:

$2.79\ \mu F$

Exercise:

Problem:

Find the total capacitance of the combination of capacitors shown in [link].



A combination of series and parallel

connections of capacitors.

Exercise:

Problem: Unreasonable Results

(a) An $8.00\ \mu F$ capacitor is connected in parallel to another capacitor, producing a total capacitance of $5.00\ \mu F$. What is the capacitance of the second capacitor? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) $-3.00\ \mu F$

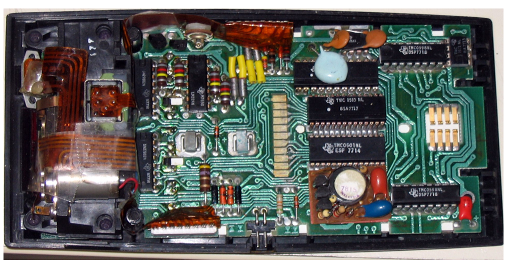
(b) You cannot have a negative value of capacitance.

(c) The assumption that the capacitors were hooked up in parallel, rather than in series, was incorrect. A parallel connection always produces a greater capacitance, while here a smaller capacitance was assumed. This could happen only if the capacitors are connected in series.

Energy Stored in Capacitors

- List some uses of capacitors.
- Express in equation form the energy stored in a capacitor.
- Explain the function of a defibrillator.

Most of us have seen dramatizations in which medical personnel use a **defibrillator** to pass an electric current through a patient's heart to get it to beat normally. (Review [link].) Often realistic in detail, the person applying the shock directs another person to "make it 400 joules this time." The energy delivered by the defibrillator is stored in a capacitor and can be adjusted to fit the situation. SI units of joules are often employed. Less dramatic is the use of capacitors in microelectronics, such as certain handheld calculators, to supply energy when batteries are charged. (See [link].) Capacitors are also used to supply energy for flash lamps on cameras.



Energy stored in the large capacitor is used to preserve the memory of an electronic calculator when its batteries are charged. (credit: Kucharek, Wikimedia Commons)

Energy stored in a capacitor is electrical potential energy, and it is thus related to the charge Q and voltage V on the capacitor. We must be careful when applying the equation for electrical potential energy $\Delta\mathrm{PE} = q\Delta V$ to a capacitor. Remember that $\Delta\mathrm{PE}$ is the potential energy of a charge q going through a voltage ΔV. But the capacitor starts with zero voltage and gradually comes up to its full voltage as it is charged. The first charge placed on a capacitor experiences a change in voltage $\Delta V = 0$, since the capacitor has zero voltage when uncharged. The final charge placed on a capacitor experiences $\Delta V = V$, since the capacitor now has its full voltage V on it. The average voltage on the capacitor during the charging process is $V/2$, and so the average voltage experienced by the full charge q is $V/2$. Thus the energy stored in a capacitor, E_{cap}, is

Equation:

$$E_{\mathrm{cap}} = \frac{QV}{2},$$

where Q is the charge on a capacitor with a voltage V applied. (Note that the energy is not QV, but $\mathrm{QV}/2$.) Charge and voltage are related to the capacitance C of a capacitor by $Q = \mathrm{CV}$, and so the expression for E_{cap} can be algebraically manipulated into three equivalent expressions:

Equation:

$$E_{\mathrm{cap}} = \frac{QV}{2} = \frac{CV^2}{2} = \frac{Q^2}{2C},$$

where Q is the charge and V the voltage on a capacitor C. The energy is in joules for a charge in coulombs, voltage in volts, and capacitance in farads.

Note:

Energy Stored in Capacitors

The energy stored in a capacitor can be expressed in three ways:

Equation:

$$E_{\text{cap}} = \frac{QV}{2} = \frac{CV^2}{2} = \frac{Q^2}{2C},$$

where Q is the charge, V is the voltage, and C is the capacitance of the capacitor. The energy is in joules for a charge in coulombs, voltage in volts, and capacitance in farads.

In a defibrillator, the delivery of a large charge in a short burst to a set of paddles across a person's chest can be a lifesaver. The person's heart attack might have arisen from the onset of fast, irregular beating of the heart—cardiac or ventricular fibrillation. The application of a large shock of electrical energy can terminate the arrhythmia and allow the body's pacemaker to resume normal patterns. Today it is common for ambulances to carry a defibrillator, which also uses an electrocardiogram to analyze the patient's heartbeat pattern. Automated external defibrillators (AED) are found in many public places ([link]). These are designed to be used by lay persons. The device automatically diagnoses the patient's heart condition and then applies the shock with appropriate energy and waveform. CPR is recommended in many cases before use of an AED.

Automated external defibrillators are found in many public places. These portable units provide verbal instructions for use in the important first few minutes for a person suffering a cardiac attack. (credit: Owain Davies, Wikimedia Commons)

Example:
Capacitance in a Heart Defibrillator

A heart defibrillator delivers 4.00×10^2 J of energy by discharging a capacitor initially at 1.00×10^4 V. What is its capacitance?

Strategy

We are given E_{cap} and V, and we are asked to find the capacitance C. Of the three expressions in the equation for E_{cap}, the most convenient relationship is

Equation:

$$E_{\text{cap}} = \frac{CV^2}{2}.$$

Solution

Solving this expression for C and entering the given values yields

Equation:

$$\begin{aligned} C &= \frac{2E_{\text{cap}}}{V^2} = \frac{2(4.00\times10^2\ \text{J})}{(1.00\times10^4\ \text{V})^2} = 8.00\times10^{-6}\ \text{F} \\ &= 8.00\ \mu\text{F}. \end{aligned}$$

Discussion
This is a fairly large, but manageable, capacitance at 1.00×10^4 V.

Section Summary

- Capacitors are used in a variety of devices, including defibrillators, microelectronics such as calculators, and flash lamps, to supply energy.
- The energy stored in a capacitor can be expressed in three ways:
  Equation:

  $$E_{\text{cap}} = \frac{\text{QV}}{2} = \frac{CV^2}{2} = \frac{Q^2}{2C},$$

  where Q is the charge, V is the voltage, and C is the capacitance of the capacitor. The energy is in joules when the charge is in coulombs, voltage is in volts, and capacitance is in farads.

Conceptual Questions

Exercise:

Problem:

How does the energy contained in a charged capacitor change when a dielectric is inserted, assuming the capacitor is isolated and its charge is constant? Does this imply that work was done?

Exercise:

Problem:

What happens to the energy stored in a capacitor connected to a battery when a dielectric is inserted? Was work done in the process?

Problems & Exercises

Exercise:

Problem:

(a) What is the energy stored in the $10.0\ \mu F$ capacitor of a heart defibrillator charged to $9.00 \times 10^3\ V$? (b) Find the amount of stored charge.

Solution:

(a) $405\ J$

(b) $90.0\ mC$

Exercise:

Problem:

In open heart surgery, a much smaller amount of energy will defibrillate the heart. (a) What voltage is applied to the $8.00\ \mu F$ capacitor of a heart defibrillator that stores 40.0 J of energy? (b) Find the amount of stored charge.

Solution:

(a) 3.16 kV

(b) 25.3 mC

Exercise:

Problem:

A $165\ \mu F$ capacitor is used in conjunction with a motor. How much energy is stored in it when 119 V is applied?

Exercise:

Problem:

Suppose you have a 9.00 V battery, a $2.00\ \mu\text{F}$ capacitor, and a $7.40\ \mu\text{F}$ capacitor. (a) Find the charge and energy stored if the capacitors are connected to the battery in series. (b) Do the same for a parallel connection.

Solution:

(a) 1.42×10^{-5} C, 6.38×10^{-5} J

(b) 8.46×10^{-5} C, 3.81×10^{-4} J

Exercise:

Problem:

A nervous physicist worries that the two metal shelves of his wood frame bookcase might obtain a high voltage if charged by static electricity, perhaps produced by friction. (a) What is the capacitance of the empty shelves if they have area $1.00 \times 10^2\ \text{m}^2$ and are 0.200 m apart? (b) What is the voltage between them if opposite charges of magnitude 2.00 nC are placed on them? (c) To show that this voltage poses a small hazard, calculate the energy stored.

Solution:

(a) 4.43×10^{-12} F

(b) 452 V

(c) 4.52×10^{-7} J

Exercise:

Problem:

Show that for a given dielectric material the maximum energy a parallel plate capacitor can store is directly proportional to the volume of dielectric ($\text{Volume} = A \cdot d$). Note that the applied voltage is limited by the dielectric strength.

Exercise:

Problem: Construct Your Own Problem

Consider a heart defibrillator similar to that discussed in [link]. Construct a problem in which you examine the charge stored in the capacitor of a defibrillator as a function of stored energy. Among the things to be considered are the applied voltage and whether it should vary with energy to be delivered, the range of energies involved, and the capacitance of the defibrillator. You may also wish to consider the much smaller energy needed for defibrillation during open-heart surgery as a variation on this problem.

Exercise:

Problem: Unreasonable Results

(a) On a particular day, it takes $9.60 \times 10^3 \text{ J}$ of electric energy to start a truck's engine. Calculate the capacitance of a capacitor that could store that amount of energy at 12.0 V. (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Solution:

(a) 133 F

(b) Such a capacitor would be too large to carry with a truck. The size of the capacitor would be enormous.

(c) It is unreasonable to assume that a capacitor can store the amount of energy needed.

Glossary

defibrillator
: a machine used to provide an electrical shock to a heart attack victim's heart in order to restore the heart's normal rhythmic pattern

Introduction to Electric Current, Resistance, and Ohm's Law class="introduction"

Electric energy in massive quantities is transmitted from this hydroelectric facility, the Srisailam power station located along the Krishna River in India, by the movement of charge—that is, by electric current. (credit: Chintohere, Wikimedia Commons)

The flicker of numbers on a handheld calculator, nerve impulses carrying signals of vision to the brain, an ultrasound device sending a signal to a computer screen, the brain sending a message for a baby to twitch its toes, an electric train pulling its load over a mountain pass, a hydroelectric plant sending energy to metropolitan and rural users—these and many other examples of electricity involve *electric current*, *the movement of charge*. Humankind has indeed harnessed electricity, the basis of technology, to improve our quality of life. Whereas the previous two chapters concentrated on static electricity and the fundamental force underlying its behavior, the next few chapters will be devoted to electric and magnetic phenomena involving current. In addition to exploring applications of electricity, we shall gain new insights into nature—in particular, the fact that all magnetism results from electric current.

Current

- Define electric current, ampere, and drift velocity
- Describe the direction of charge flow in conventional current.
- Use drift velocity to calculate current and vice versa.

Electric Current

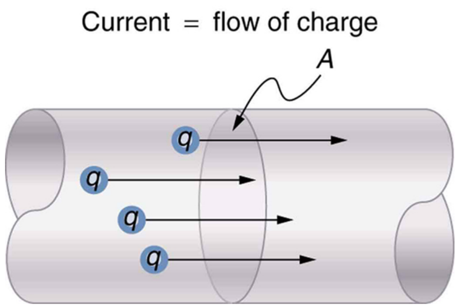
Electric current is defined to be the rate at which charge flows. A large current, such as that used to start a truck engine, moves a large amount of charge in a small time, whereas a small current, such as that used to operate a hand-held calculator, moves a small amount of charge over a long period of time. In equation form, **electric current** I is defined to be
Equation:

$$I = \frac{\Delta Q}{\Delta t},$$

where ΔQ is the amount of charge passing through a given area in time Δt. (As in previous chapters, initial time is often taken to be zero, in which case $\Delta t = t$.) (See [link].) The SI unit for current is the **ampere** (A), named for the French physicist André-Marie Ampère (1775–1836). Since $I = \Delta Q/\Delta t$, we see that an ampere is one coulomb per second:
Equation:

$$1\ \text{A} = 1\ \text{C/s}$$

Not only are fuses and circuit breakers rated in amperes (or amps), so are many electrical appliances.



The rate of flow of charge is current. An ampere is the flow of one coulomb through an area in one second.

Example:
Calculating Currents: Current in a Truck Battery and a Handheld Calculator
(a) What is the current involved when a truck battery sets in motion 720 C of charge in 4.00 s while starting an engine? (b) How long does it take 1.00 C of charge to flow through a handheld calculator if a 0.300-mA current is flowing?
Strategy
We can use the definition of current in the equation $I = \Delta Q/\Delta t$ to find the current in part (a), since charge and time are given. In part (b), we rearrange the definition of current and use the given values of charge and current to find the time required.
Solution for (a)
Entering the given values for charge and time into the definition of current gives
Equation:

$$\begin{aligned} I &= \frac{\Delta Q}{\Delta t} = \frac{720\text{ C}}{4.00\text{ s}} = 180\text{ C/s} \\ &= 180\text{ A.} \end{aligned}$$

Discussion for (a)

This large value for current illustrates the fact that a large charge is moved in a small amount of time. The currents in these “starter motors” are fairly large because large frictional forces need to be overcome when setting something in motion.

Solution for (b)

Solving the relationship $I=\Delta Q/\Delta t$ for time Δt, and entering the known values for charge and current gives

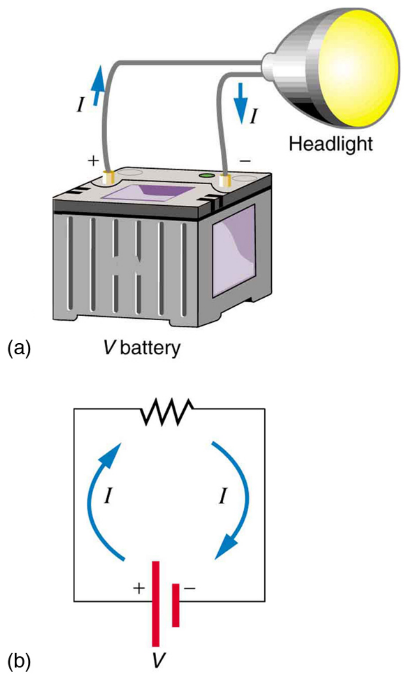
Equation:

$$\begin{aligned}\Delta t &= \frac{\Delta Q}{I}=\frac{1.00\text{ C}}{0.300\times 10^{-3}\text{ C/s}}\\ &= 3.33\times 10^{3}\text{ s.}\end{aligned}$$

Discussion for (b)

This time is slightly less than an hour. The small current used by the hand-held calculator takes a much longer time to move a smaller charge than the large current of the truck starter. So why can we operate our calculators only seconds after turning them on? It’s because calculators require very little energy. Such small current and energy demands allow handheld calculators to operate from solar cells or to get many hours of use out of small batteries. Remember, calculators do not have moving parts in the same way that a truck engine has with cylinders and pistons, so the technology requires smaller currents.

[link] shows a simple circuit and the standard schematic representation of a battery, conducting path, and load (a resistor). Schematics are very useful in visualizing the main features of a circuit. A single schematic can represent a wide variety of situations. The schematic in [link] (b), for example, can represent anything from a truck battery connected to a headlight lighting the street in front of the truck to a small battery connected to a penlight lighting a keyhole in a door. Such schematics are useful because the analysis is the same for a wide variety of situations. We need to understand a few schematics to apply the concepts and analysis to many more situations.



(a) A simple electric circuit. A closed path for current to flow through is supplied by conducting wires connecting a load to the terminals of a battery. (b) In this schematic, the battery is represented by the two parallel red lines, conducting wires are shown as straight lines, and the zigzag represents the load. The schematic represents a wide

variety of similar circuits.

Note that the direction of current flow in [link] is from positive to negative. *The direction of conventional current is the direction that positive charge would flow*. Depending on the situation, positive charges, negative charges, or both may move. In metal wires, for example, current is carried by electrons—that is, negative charges move. In ionic solutions, such as salt water, both positive and negative charges move. This is also true in nerve cells. A Van de Graaff generator used for nuclear research can produce a current of pure positive charges, such as protons. [link] illustrates the movement of charged particles that compose a current. The fact that conventional current is taken to be in the direction that positive charge would flow can be traced back to American politician and scientist Benjamin Franklin in the 1700s. He named the type of charge associated with electrons negative, long before they were known to carry current in so many situations. Franklin, in fact, was totally unaware of the small-scale structure of electricity.

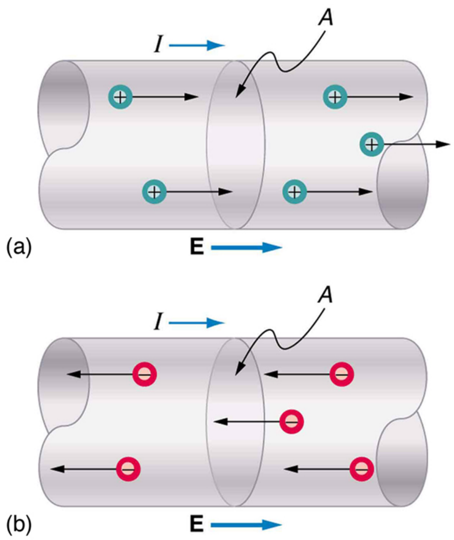
It is important to realize that there is an electric field in conductors responsible for producing the current, as illustrated in [link]. Unlike static electricity, where a conductor in equilibrium cannot have an electric field in it, conductors carrying a current have an electric field and are not in static equilibrium. An electric field is needed to supply energy to move the charges.

Note:

Making Connections: Take-Home Investigation—Electric Current Illustration

Find a straw and little peas that can move freely in the straw. Place the straw flat on a table and fill the straw with peas. When you pop one pea in at one end, a different pea should pop out the other end. This demonstration is an analogy for an electric current. Identify what compares

to the electrons and what compares to the supply of energy. What other analogies can you find for an electric current?
Note that the flow of peas is based on the peas physically bumping into each other; electrons flow due to mutually repulsive electrostatic forces.



Current I is the rate at which charge moves through an area A, such as the cross-section of a wire. Conventional current is defined to move in the direction of the electric field. (a) Positive charges move in the direction of the electric field and the same direction as conventional current. (b) Negative charges move in the direction opposite to the electric field. Conventional

current is in the direction opposite to the movement of negative charge. The flow of electrons is sometimes referred to as electronic flow.

Example:

Calculating the Number of Electrons that Move through a Calculator

If the 0.300-mA current through the calculator mentioned in the [link] example is carried by electrons, how many electrons per second pass through it?

Strategy

The current calculated in the previous example was defined for the flow of positive charge. For electrons, the magnitude is the same, but the sign is opposite, $I_{\text{electrons}} = -0.300 \times 10^{-3}\ \text{C/s}$ .Since each electron (e^-) has a charge of -1.60×10^{-19} C, we can convert the current in coulombs per second to electrons per second.

Solution

Starting with the definition of current, we have

Equation:

$$I_{\text{electrons}} = \frac{\Delta Q_{\text{electrons}}}{\Delta t} = \frac{-0.300 \times 10^{-3}\ \text{C}}{\text{s}}.$$

We divide this by the charge per electron, so that

Equation:

$$\begin{aligned} \frac{e^-}{\text{s}} &= \frac{-0.300\times10^{-3}\ \text{C}}{\text{s}} \times \frac{1\ e^-}{-1.60\times10^{-19}\ \text{C}} \\ &= 1.88\times10^{15}\ \frac{e^-}{\text{s}}. \end{aligned}$$

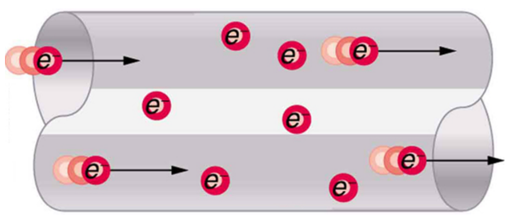
Discussion

There are so many charged particles moving, even in small currents, that individual charges are not noticed, just as individual water molecules are not noticed in water flow. Even more amazing is that they do not always keep moving forward like soldiers in a parade. Rather they are like a crowd of people with movement in different directions but a general trend to move forward. There are lots of collisions with atoms in the metal wire and, of course, with other electrons.

Drift Velocity

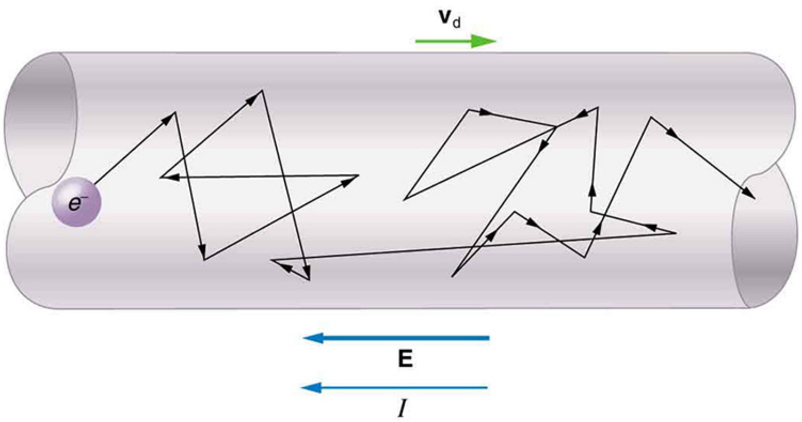
Electrical signals are known to move very rapidly. Telephone conversations carried by currents in wires cover large distances without noticeable delays. Lights come on as soon as a switch is flicked. Most electrical signals carried by currents travel at speeds on the order of $10^8\ \mathrm{m/s}$, a significant fraction of the speed of light. Interestingly, the individual charges that make up the current move *much* more slowly on average, typically drifting at speeds on the order of $10^{-4}\ \mathrm{m/s}$. How do we reconcile these two speeds, and what does it tell us about standard conductors?

The high speed of electrical signals results from the fact that the force between charges acts rapidly at a distance. Thus, when a free charge is forced into a wire, as in [link], the incoming charge pushes other charges ahead of it, which in turn push on charges farther down the line. The density of charge in a system cannot easily be increased, and so the signal is passed on rapidly. The resulting electrical shock wave moves through the system at nearly the speed of light. To be precise, this rapidly moving signal or shock wave is a rapidly propagating change in electric field.



When charged particles are forced into this volume of a conductor, an equal number are quickly forced to leave. The repulsion between like charges makes it difficult to increase the number of charges in a volume. Thus, as one charge enters, another leaves almost immediately, carrying the signal rapidly forward.

Good conductors have large numbers of free charges in them. In metals, the free charges are free electrons. [link] shows how free electrons move through an ordinary conductor. The distance that an individual electron can move between collisions with atoms or other electrons is quite small. The electron paths thus appear nearly random, like the motion of atoms in a gas. But there is an electric field in the conductor that causes the electrons to drift in the direction shown (opposite to the field, since they are negative). The **drift velocity** v_{d} is the average velocity of the free charges. Drift velocity is quite small, since there are so many free charges. If we have an estimate of the density of free electrons in a conductor, we can calculate the drift velocity for a given current. The larger the density, the lower the velocity required for a given current.



Free electrons moving in a conductor make many collisions with other electrons and atoms. The path of one electron is shown. The average velocity of the free charges is called the drift velocity, v_d, and it is in the direction opposite to the electric field for electrons. The collisions normally transfer energy to the conductor, requiring a constant supply of energy to maintain a steady current.

Note:
Conduction of Electricity and Heat
Good electrical conductors are often good heat conductors, too. This is because large numbers of free electrons can carry electrical current and can transport thermal energy.

The free-electron collisions transfer energy to the atoms of the conductor. The electric field does work in moving the electrons through a distance, but that work does not increase the kinetic energy (nor speed, therefore) of the electrons. The work is transferred to the conductor’s atoms, possibly

increasing temperature. Thus a continuous power input is required to keep a current flowing. An exception, of course, is found in superconductors, for reasons we shall explore in a later chapter. Superconductors can have a steady current without a continual supply of energy—a great energy savings. In contrast, the supply of energy can be useful, such as in a lightbulb filament. The supply of energy is necessary to increase the temperature of the tungsten filament, so that the filament glows.

Note:
Making Connections: Take-Home Investigation—Filament Observations
Find a lightbulb with a filament. Look carefully at the filament and describe its structure. To what points is the filament connected?

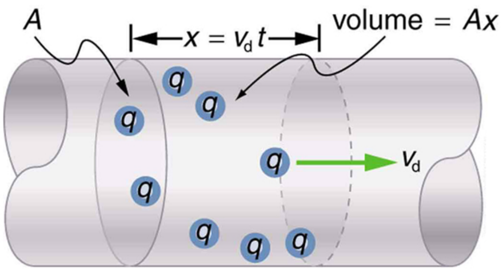
We can obtain an expression for the relationship between current and drift velocity by considering the number of free charges in a segment of wire, as illustrated in [link]. *The number of free charges per unit volume* is given the symbol n and depends on the material. The shaded segment has a volume Ax, so that the number of free charges in it is nAx. The charge ΔQ in this segment is thus qnAx, where q is the amount of charge on each carrier. (Recall that for electrons, q is $-1.60 \times 10^{-19}\ \mathrm{C}$.) Current is charge moved per unit time; thus, if all the original charges move out of this segment in time Δt, the current is
Equation:

$$I = \frac{\Delta Q}{\Delta t} = \frac{\mathrm{qnAx}}{\Delta t}.$$

Note that $x/\Delta t$ is the magnitude of the drift velocity, v_{d}, since the charges move an average distance x in a time Δt. Rearranging terms gives
Equation:

$$I = \mathrm{nqAv}_{\mathrm{d}},$$

where I is the current through a wire of cross-sectional area A made of a material with a free charge density n. The carriers of the current each have charge q and move with a drift velocity of magnitude v_d.



All the charges in the shaded volume of this wire move out in a time t, having a drift velocity of magnitude $v_d = x/t$. See text for further discussion.

Note that simple drift velocity is not the entire story. The speed of an electron is much greater than its drift velocity. In addition, not all of the electrons in a conductor can move freely, and those that do might move somewhat faster or slower than the drift velocity. So what do we mean by free electrons? Atoms in a metallic conductor are packed in the form of a lattice structure. Some electrons are far enough away from the atomic nuclei that they do not experience the attraction of the nuclei as much as the inner electrons do. These are the free electrons. They are not bound to a single atom but can instead move freely among the atoms in a "sea" of electrons. These free electrons respond by accelerating when an electric field is applied. Of course as they move they collide with the atoms in the lattice and other electrons, generating thermal energy, and the conductor gets warmer. In an insulator, the organization of the atoms and the structure do not allow for such free electrons.

Example:

Calculating Drift Velocity in a Common Wire

Calculate the drift velocity of electrons in a 12-gauge copper wire (which has a diameter of 2.053 mm) carrying a 20.0-A current, given that there is one free electron per copper atom. (Household wiring often contains 12-gauge copper wire, and the maximum current allowed in such wire is usually 20 A.) The density of copper is $8.80 \times 10^3 \text{ kg/m}^3$.

Strategy

We can calculate the drift velocity using the equation $I = nqAv_{\text{d}}$. The current $I = 20.0 \text{ A}$ is given, and $q = -1.60 \times 10^{-19}\text{C}$ is the charge of an electron. We can calculate the area of a cross-section of the wire using the formula $A = \pi r^2$, where r is one-half the given diameter, 2.053 mm. We are given the density of copper, $8.80 \times 10^3 \text{ kg/m}^3$, and the periodic table shows that the atomic mass of copper is 63.54 g/mol. We can use these two quantities along with Avogadro's number, $6.02 \times 10^{23} \text{ atoms/mol}$, to determine n, the number of free electrons per cubic meter.

Solution

First, calculate the density of free electrons in copper. There is one free electron per copper atom. Therefore, is the same as the number of copper atoms per m^3. We can now find n as follows:

Equation:

$$\begin{aligned} n &= \frac{1\ e^-}{\text{atom}} \times \frac{6.02\times10^{23}\text{ atoms}}{\text{mol}} \times \frac{1\text{ mol}}{63.54\text{ g}} \times \frac{1000\text{ g}}{\text{kg}} \times \frac{8.80\times10^{3}\text{ kg}}{1\text{ m}^3} \\ &= 8.342 \times 10^{28}\ e^-/\text{m}^3. \end{aligned}$$

The cross-sectional area of the wire is

Equation:

$$\begin{aligned} A &= \pi r^2 \\ &= \pi \left(\frac{2.053\times10^{-3}\text{ m}}{2}\right)^2 \\ &= 3.310 \times 10^{-6}\text{ m}^2. \end{aligned}$$

Rearranging $I = nqAv_{\text{d}}$ to isolate drift velocity gives

Equation:

$$v_d = \frac{I}{nqA}$$

$$= \frac{20.0\ \text{A}}{(8.342\times10^{28}/\text{m}^3)(-1.60\times10^{-19}\ \text{C})(3.310\times10^{-6}\ \text{m}^2)}$$

$$= -4.53 \times 10^{-4}\ \text{m/s}.$$

Discussion

The minus sign indicates that the negative charges are moving in the direction opposite to conventional current. The small value for drift velocity (on the order of 10^{-4} m/s) confirms that the signal moves on the order of 10^{12} times faster (about 10^{8} m/s) than the charges that carry it.

Section Summary

- Electric current I is the rate at which charge flows, given by **Equation:**

$$I = \frac{\Delta Q}{\Delta t},$$

  where ΔQ is the amount of charge passing through an area in time Δt.
- The direction of conventional current is taken as the direction in which positive charge moves.
- The SI unit for current is the ampere (A), where $1\ \text{A} = 1\ \text{C/s}$.
- Current is the flow of free charges, such as electrons and ions.
- Drift velocity v_d is the average speed at which these charges move.
- Current I is proportional to drift velocity v_d, as expressed in the relationship $I = \text{nqAv}_d$. Here, I is the current through a wire of cross-sectional area A. The wire's material has a free-charge density n, and each carrier has charge q and a drift velocity v_d.
- Electrical signals travel at speeds about 10^{12} times greater than the drift velocity of free electrons.

Conceptual Questions

Exercise:

Problem:

Can a wire carry a current and still be neutral—that is, have a total charge of zero? Explain.

Exercise:

Problem:

Car batteries are rated in ampere-hours ($\mathrm{A \cdot h}$). To what physical quantity do ampere-hours correspond (voltage, charge, . . .), and what relationship do ampere-hours have to energy content?

Exercise:

Problem:

If two different wires having identical cross-sectional areas carry the same current, will the drift velocity be higher or lower in the better conductor? Explain in terms of the equation $v_{\mathrm{d}} = \frac{I}{\mathrm{nqA}}$, by considering how the density of charge carriers n relates to whether or not a material is a good conductor.

Exercise:

Problem:

Why are two conducting paths from a voltage source to an electrical device needed to operate the device?

Exercise:

Problem:

In cars, one battery terminal is connected to the metal body. How does this allow a single wire to supply current to electrical devices rather than two wires?

Exercise:

Problem:

Why isn’t a bird sitting on a high-voltage power line electrocuted? Contrast this with the situation in which a large bird hits two wires simultaneously with its wings.

Problems & Exercises

Exercise:

Problem:

What is the current in milliamperes produced by the solar cells of a pocket calculator through which 4.00 C of charge passes in 4.00 h?

Solution:

0.278 mA

Exercise:

Problem:

A total of 600 C of charge passes through a flashlight in 0.500 h. What is the average current?

Exercise:

Problem:

What is the current when a typical static charge of $0.250\ \mu\text{C}$ moves from your finger to a metal doorknob in $1.00\ \mu\text{s}$?

Solution:

0.250 A

Exercise:

Problem:

Find the current when 2.00 nC jumps between your comb and hair over a 0.500 - μs time interval.

Exercise:

Problem:

A large lightning bolt had a 20,000-A current and moved 30.0 C of charge. What was its duration?

Solution:

1.50ms

Exercise:

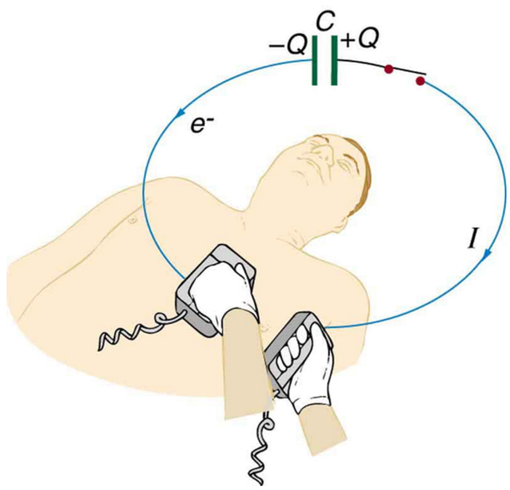
Problem:

The 200-A current through a spark plug moves 0.300 mC of charge. How long does the spark last?

Exercise:

Problem:

(a) A defibrillator sends a 6.00-A current through the chest of a patient by applying a 10,000-V potential as in the figure below. What is the resistance of the path? (b) The defibrillator paddles make contact with the patient through a conducting gel that greatly reduces the path resistance. Discuss the difficulties that would ensue if a larger voltage were used to produce the same current through the patient, but with the path having perhaps 50 times the resistance. (Hint: The current must be about the same, so a higher voltage would imply greater power. Use this equation for power: $P = I^2R$.)



The capacitor in a defibrillation unit drives a current through the heart of a patient.

Solution:

(a) $1.67\text{k}\Omega$

(b) If a 50 times larger resistance existed, keeping the current about the same, the power would be increased by a factor of about 50 (based on the equation $P = I^2R$), causing much more energy to be transferred to the skin, which could cause serious burns. The gel used reduces the resistance, and therefore reduces the power transferred to the skin.

Exercise:

Problem:

During open-heart surgery, a defibrillator can be used to bring a patient out of cardiac arrest. The resistance of the path is $500\ \Omega$ and a 10.0-mA current is needed. What voltage should be applied?

Exercise:

Problem:

(a) A defibrillator passes 12.0 A of current through the torso of a person for 0.0100 s. How much charge moves? (b) How many electrons pass through the wires connected to the patient? (See figure two problems earlier.)

Solution:

(a) 0.120 C

(b) $7.50 \times 10^{17} \text{electrons}$

Exercise:

Problem:

A clock battery wears out after moving 10,000 C of charge through the clock at a rate of 0.500 mA. (a) How long did the clock run? (b) How many electrons per second flowed?

Exercise:

Problem:

The batteries of a submerged non-nuclear submarine supply 1000 A at full speed ahead. How long does it take to move Avogadro's number (6.02×10^{23}) of electrons at this rate?

Solution:

96.3 s

Exercise:

Problem:

Electron guns are used in X-ray tubes. The electrons are accelerated through a relatively large voltage and directed onto a metal target, producing X-rays. (a) How many electrons per second strike the target if the current is 0.500 mA? (b) What charge strikes the target in 0.750 s?

Exercise:

Problem:

A large cyclotron directs a beam of He^{++} nuclei onto a target with a beam current of 0.250 mA. (a) How many He^{++} nuclei per second is this? (b) How long does it take for 1.00 C to strike the target? (c) How long before 1.00 mol of He^{++} nuclei strike the target?

Solution:

(a) $7.81 \times 10^{14}\ \mathrm{He}^{++}\ \mathrm{nuclei/s}$

(b) $4.00 \times 10^{3}\ \mathrm{s}$

(c) $7.71 \times 10^{8}\ \mathrm{s}$

Exercise:

Problem:

Repeat the above example on [link], but for a wire made of silver and given there is one free electron per silver atom.

Exercise:

Problem:

Using the results of the above example on [link], find the drift velocity in a copper wire of twice the diameter and carrying 20.0 A.

Solution:

$-1.13 \times 10^{-4} \mathrm{m/s}$

Exercise:

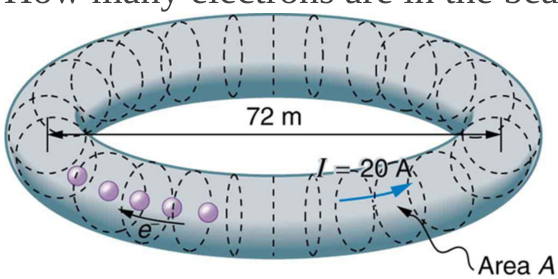
Problem:

A 14-gauge copper wire has a diameter of 1.628 mm. What magnitude current flows when the drift velocity is 1.00 mm/s? (See above example on [link] for useful information.)

Exercise:

Problem:

SPEAR, a storage ring about 72.0 m in diameter at the Stanford Linear Accelerator (closed in 2009), has a 20.0-A circulating beam of electrons that are moving at nearly the speed of light. (See [link].) How many electrons are in the beam?



Electrons circulating in the storage ring called SPEAR constitute a 20.0-A current. Because they travel close to the speed of light, each electron completes many orbits in each second.

Solution:

$9.42 \times 10^{13} \text{electrons}$

Glossary

electric current
: the rate at which charge flows, $I = \Delta Q/\Delta t$

ampere
: (amp) the SI unit for current; 1 A = 1 C/s

drift velocity
: the average velocity at which free charges flow in response to an electric field

Ohm’s Law: Resistance and Simple Circuits

- Explain the origin of Ohm’s law.
- Calculate voltages, currents, or resistances with Ohm’s law.
- Explain what an ohmic material is.
- Describe a simple circuit.

What drives current? We can think of various devices—such as batteries, generators, wall outlets, and so on—which are necessary to maintain a current. All such devices create a potential difference and are loosely referred to as voltage sources. When a voltage source is connected to a conductor, it applies a potential difference that creates an electric field. The electric field in turn exerts force on charges, causing current.

Ohm’s Law

The current that flows through most substances is directly proportional to the voltage applied to it. The German physicist Georg Simon Ohm (1787–1854) was the first to demonstrate experimentally that the current in a metal wire is *directly proportional to the voltage applied*:
Equation:

This important relationship is known as **Ohm’s law**. It can be viewed as a cause-and-effect relationship, with voltage the cause and current the effect. This is an empirical law like that for friction—an experimentally observed phenomenon. Such a linear relationship doesn’t always occur.

Resistance and Simple Circuits

If voltage drives current, what impedes it? The electric property that impedes current (crudely similar to friction and air resistance) is called **resistance** . Collisions of moving charges with atoms and molecules in a substance transfer energy to the substance and limit current. Resistance is defined as inversely proportional to current, or

Equation:

—

Thus, for example, current is cut in half if resistance doubles. Combining the relationships of current to voltage and current to resistance gives

Equation:

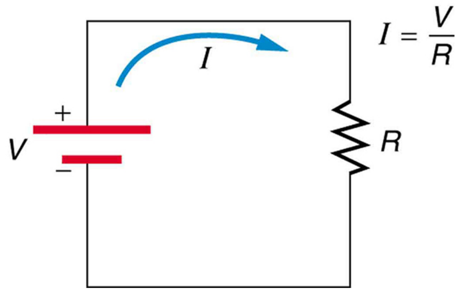
—

This relationship is also called Ohm’s law. Ohm’s law in this form really defines resistance for certain materials. Ohm’s law (like Hooke’s law) is not universally valid. The many substances for which Ohm’s law holds are called **ohmic**. These include good conductors like copper and aluminum, and some poor conductors under certain circumstances. Ohmic materials have a resistance that is independent of voltage and current . An object that has simple resistance is called a *resistor*, even if its resistance is small. The unit for resistance is an **ohm** and is given the symbol (upper case Greek omega). Rearranging gives , and so the units of resistance are 1 ohm = 1 volt per ampere:

Equation:

—

[link] shows the schematic for a simple circuit. A **simple circuit** has a single voltage source and a single resistor. The wires connecting the voltage source to the resistor can be assumed to have negligible resistance, or their resistance can be included in .



A simple electric circuit in which a closed path for current to flow is supplied by conductors (usually metal wires) connecting a load to the terminals of a battery, represented by the red parallel lines. The zigzag symbol represents the single resistor and includes any resistance in the connections to the voltage source.

Example:

Calculating Resistance: An Automobile Headlight

What is the resistance of an automobile headlight through which 2.50 A flows when 12.0 V is applied to it?

Strategy

We can rearrange Ohm’s law as stated by and use it to find the resistance.

Solution

Rearranging and substituting known values gives

Equation:

Discussion

This is a relatively small resistance, but it is larger than the cold resistance of the headlight. As we shall see in [Resistance and Resistivity], resistance usually increases with temperature, and so the bulb has a lower resistance when it is first switched on and will draw considerably more current during its brief warm-up period.

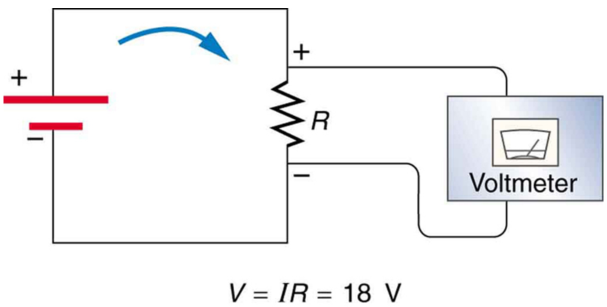
Resistances range over many orders of magnitude. Some ceramic insulators, such as those used to support power lines, have resistances of or more. A dry person may have a hand-to-foot resistance of , whereas the resistance of the human heart is about . A meter-long piece of large-diameter copper wire may have a resistance of , and superconductors have no resistance at all (they are non-ohmic). Resistance is related to the shape of an object and the material of which it is composed, as will be seen in [Resistance and Resistivity].

Additional insight is gained by solving for yielding

Equation:

This expression for can be interpreted as the *voltage drop across a resistor produced by the flow of current* . The phrase *drop* is often used for this voltage. For instance, the headlight in [link] has an drop of 12.0 V. If voltage is measured at various points in a circuit, it will be seen to increase at the voltage source and decrease at the resistor. Voltage is similar to fluid pressure. The voltage source is like a pump, creating a pressure difference, causing current—the flow of charge. The resistor is like a pipe that reduces pressure and limits flow because of its resistance. Conservation of energy has important consequences here. The voltage source supplies

energy (causing an electric field and a current), and the resistor converts it to another form (such as thermal energy). In a simple circuit (one with a single simple resistor), the voltage supplied by the source equals the voltage drop across the resistor, since , and the same flows through each. Thus the energy supplied by the voltage source and the energy converted by the resistor are equal. (See [link].)



The voltage drop across a resistor in a simple circuit equals the voltage output of the battery.

Note:
Making Connections: Conservation of Energy
In a simple electrical circuit, the sole resistor converts energy supplied by the source into another form. Conservation of energy is evidenced here by the fact that all of the energy supplied by the source is converted to another form by the resistor alone. We will find that conservation of energy has other important applications in circuits and is a powerful tool in circuit analysis.

Note:

PhET Explorations: Ohm's Law
See how the equation form of Ohm's law relates to a simple circuit. Adjust the voltage and resistance, and see the current change according to Ohm's law. The sizes of the symbols in the equation change to match the circuit diagram.
https://phet.colorado.edu/sims/html/ohms-law/latest/ohms-law_en.html

Section Summary

- A simple circuit *is* one in which there is a single voltage source and a single resistance.
- One statement of Ohm's law gives the relationship between current , voltage , and resistance in a simple circuit to be —
- Resistance has units of ohms ( ), related to volts and amperes by .
- There is a voltage or drop across a resistor, caused by the current flowing through it, given by .

Conceptual Questions

Exercise:

Problem:

The drop across a resistor means that there is a change in potential or voltage across the resistor. Is there any change in current as it passes through a resistor? Explain.

Exercise:

Problem:

How is the drop in a resistor similar to the pressure drop in a fluid flowing through a pipe?

Problems & Exercises

Exercise:

Problem:

What current flows through the bulb of a 3.00-V flashlight when its hot resistance is ?

Solution:

0.833 A

Exercise:

Problem:

Calculate the effective resistance of a pocket calculator that has a 1.35-V battery and through which 0.200 mA flows.

Exercise:

Problem:

What is the effective resistance of a car's starter motor when 150 A flows through it as the car battery applies 11.0 V to the motor?

Solution:

Exercise:

Problem:

How many volts are supplied to operate an indicator light on a DVD player that has a resistance of , given that 25.0 mA passes through it?

Exercise:

Problem:

(a) Find the voltage drop in an extension cord having a resistance and through which 5.00 A is flowing. (b) A cheaper cord utilizes thinner wire and has a resistance of . What is the voltage drop in it when 5.00 A flows? (c) Why is the voltage to whatever appliance is being used reduced by this amount? What is the effect on the appliance?

Solution:

(a) 0.300 V

(b) 1.50 V

(c) The voltage supplied to whatever appliance is being used is reduced because the total voltage drop from the wall to the final output of the appliance is fixed. Thus, if the voltage drop across the extension cord is large, the voltage drop across the appliance is significantly decreased, so the power output by the appliance can be significantly decreased, reducing the ability of the appliance to work properly.

Exercise:

Problem:

A power transmission line is hung from metal towers with glass insulators having a resistance of What current flows through the insulator if the voltage is 200 kV? (Some high-voltage lines are DC.)

Glossary

Ohm's law
: an empirical relation stating that the current I is proportional to the potential difference V, $\propto V$; it is often written as $I = V/R$, where R is the resistance

resistance

the electric property that impedes current; for ohmic materials, it is the ratio of voltage to current, $R = V/I$

ohm

the unit of resistance, given by $1\Omega = 1$ V/A

ohmic

a type of a material for which Ohm's law is valid

simple circuit

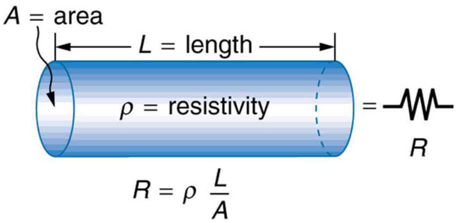
a circuit with a single voltage source and a single resistor

Resistance and Resistivity

- Explain the concept of resistivity.
- Use resistivity to calculate the resistance of specified configurations of material.
- Use the thermal coefficient of resistivity to calculate the change of resistance with temperature.

Material and Shape Dependence of Resistance

The resistance of an object depends on its shape and the material of which it is composed. The cylindrical resistor in [link] is easy to analyze, and, by so doing, we can gain insight into the resistance of more complicated shapes. As you might expect, the cylinder's electric resistance R is directly proportional to its length L, similar to the resistance of a pipe to fluid flow. The longer the cylinder, the more collisions charges will make with its atoms. The greater the diameter of the cylinder, the more current it can carry (again similar to the flow of fluid through a pipe). In fact, R is inversely proportional to the cylinder's cross-sectional area A.



A uniform cylinder of length L and cross-sectional area A. Its resistance to the flow of current is similar to the resistance posed by a pipe to fluid flow. The longer the cylinder, the greater its

resistance. The larger its cross-sectional area A, the smaller its resistance.

For a given shape, the resistance depends on the material of which the object is composed. Different materials offer different resistance to the flow of charge. We define the **resistivity** ρ of a substance so that the **resistance** R of an object is directly proportional to ρ. Resistivity ρ is an *intrinsic* property of a material, independent of its shape or size. The resistance R of a uniform cylinder of length L, of cross-sectional area A, and made of a material with resistivity ρ, is

Equation:

$$R = \frac{\rho L}{A}.$$

[link] gives representative values of ρ. The materials listed in the table are separated into categories of conductors, semiconductors, and insulators, based on broad groupings of resistivities. Conductors have the smallest resistivities, and insulators have the largest; semiconductors have intermediate resistivities. Conductors have varying but large free charge densities, whereas most charges in insulators are bound to atoms and are not free to move. Semiconductors are intermediate, having far fewer free charges than conductors, but having properties that make the number of free charges depend strongly on the type and amount of impurities in the semiconductor. These unique properties of semiconductors are put to use in modern electronics, as will be explored in later chapters.

Material	**Resistivity ρ ($\Omega \cdot \mathrm{m}$)**
Conductors	
Silver	1.59×10^{-8}
Copper	1.72×10^{-8}
Gold	2.44×10^{-8}
Aluminum	2.65×10^{-8}
Tungsten	5.6×10^{-8}
Iron	9.71×10^{-8}
Platinum	10.6×10^{-8}
Steel	20×10^{-8}
Lead	22×10^{-8}

Material	**Resistivity ρ ($\Omega \cdot \mathrm{m}$)**
Manganin (Cu, Mn, Ni alloy)	44×10^{-8}
Constantan (Cu, Ni alloy)	49×10^{-8}
Mercury	96×10^{-8}
Nichrome (Ni, Fe, Cr alloy)	100×10^{-8}
Semiconductors[footnote] Values depend strongly on amounts and types of impurities	
Carbon (pure)	3.5×10^{-5}
Carbon	$(3.5 - 60) \times 10^{-5}$
Germanium (pure)	600×10^{-3}
Germanium	$(1 - 600) \times 10^{-3}$

Material	**Resistivity ρ ($\Omega \cdot \mathrm{m}$)**
Silicon (pure)	2300
Silicon	0.1–2300
Insulators	
Amber	5×10^{14}
Glass	$10^9 - 10^{14}$
Lucite	$>10^{13}$
Mica	$10^{11} - 10^{15}$
Quartz (fused)	75×10^{16}
Rubber (hard)	$10^{13} - 10^{16}$
Sulfur	10^{15}

Material	Resistivity ρ ($\Omega \cdot \mathrm{m}$)
Teflon	$>10^{13}$
Wood	$10^{8}-10^{11}$

Resistivities ρ of Various materials at $20^{\circ}\mathrm{C}$

Example:
Calculating Resistor Diameter: A Headlight Filament
A car headlight filament is made of tungsten and has a cold resistance of $0.350\ \Omega$. If the filament is a cylinder 4.00 cm long (it may be coiled to save space), what is its diameter?
Strategy
We can rearrange the equation $R=\frac{\rho L}{A}$ to find the cross-sectional area A of the filament from the given information. Then its diameter can be found by assuming it has a circular cross-section.
Solution
The cross-sectional area, found by rearranging the expression for the resistance of a cylinder given in $R=\frac{\rho L}{A}$, is
Equation:

$$A=\frac{\rho L}{R}.$$

Substituting the given values, and taking ρ from [link], yields
Equation:

$$\begin{aligned} A &= \frac{(5.6\times10^{-8}\ \Omega\cdot\mathrm{m})(4.00\times10^{-2}\ \mathrm{m})}{0.350\ \Omega} \\ &= 6.40\times10^{-9}\ \mathrm{m}^{2}. \end{aligned}$$

The area of a circle is related to its diameter D by
Equation:

$$A = \frac{\pi D^2}{4}.$$

Solving for the diameter D, and substituting the value found for A, gives
Equation:

$$\begin{aligned} D &= 2\left(\frac{A}{p}\right)^{\frac{1}{2}} = 2\left(\frac{6.40 \times 10^{-9}\ \text{m}^2}{3.14}\right)^{\frac{1}{2}} \\ &= 9.0 \times 10^{-5}\ \text{m}. \end{aligned}$$

Discussion
The diameter is just under a tenth of a millimeter. It is quoted to only two digits, because ρ is known to only two digits.

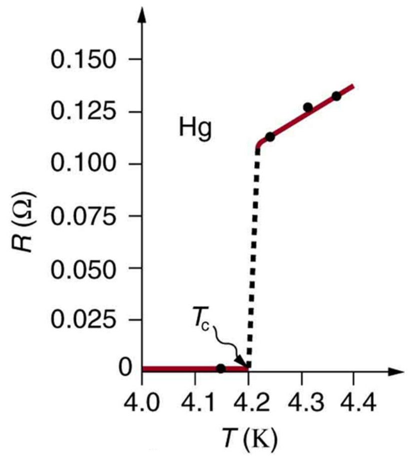
Temperature Variation of Resistance

The resistivity of all materials depends on temperature. Some even become superconductors (zero resistivity) at very low temperatures. (See [link].) Conversely, the resistivity of conductors increases with increasing temperature. Since the atoms vibrate more rapidly and over larger distances at higher temperatures, the electrons moving through a metal make more collisions, effectively making the resistivity higher. Over relatively small temperature changes (about 100ºC or less), resistivity ρ varies with temperature change ΔT as expressed in the following equation
Equation:

$$\rho = \rho_0 (1 + \alpha \Delta T),$$

where ρ_0 is the original resistivity and α is the **temperature coefficient of resistivity**. (See the values of α in [link] below.) For larger temperature changes, α may vary or a nonlinear equation may be needed to find ρ. Note

that α is positive for metals, meaning their resistivity increases with temperature. Some alloys have been developed specifically to have a small temperature dependence. Manganin (which is made of copper, manganese and nickel), for example, has α close to zero (to three digits on the scale in [link]), and so its resistivity varies only slightly with temperature. This is useful for making a temperature-independent resistance standard, for example.



The resistance of a sample of mercury is zero at very low temperatures—it is a superconductor up to about 4.2 K. Above that critical temperature, its resistance makes a sudden jump and then increases nearly linearly with temperature.

Material	Coefficient α(1/°C)[footnote] Values at 20°C.
Conductors	
Silver	3.8×10^{-3}
Copper	3.9×10^{-3}
Gold	3.4×10^{-3}
Aluminum	3.9×10^{-3}
Tungsten	4.5×10^{-3}
Iron	5.0×10^{-3}
Platinum	3.93×10^{-3}
Lead	3.9×10^{-3}
Manganin (Cu, Mn, Ni alloy)	0.000×10^{-3}

Material	**Coefficient α(1/°C)[footnote] Values at 20°C.**
Constantan (Cu, Ni alloy)	0.002×10^{-3}
Mercury	0.89×10^{-3}
Nichrome (Ni, Fe, Cr alloy)	0.4×10^{-3}
Semiconductors	
Carbon (pure)	-0.5×10^{-3}
Germanium (pure)	-50×10^{-3}
Silicon (pure)	-70×10^{-3}

Tempature Coefficients of Resistivity α

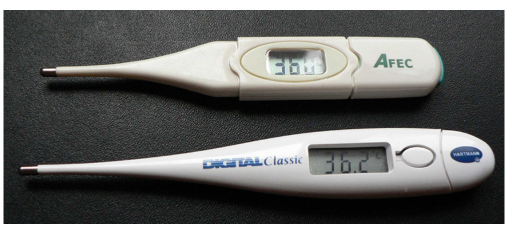
Note also that α is negative for the semiconductors listed in [link], meaning that their resistivity decreases with increasing temperature. They become better conductors at higher temperature, because increased thermal agitation increases the number of free charges available to carry current. This property of decreasing ρ with temperature is also related to the type and amount of impurities present in the semiconductors.

The resistance of an object also depends on temperature, since R_0 is directly proportional to ρ. For a cylinder we know $R = \rho L/A$, and so, if L and A do not change greatly with temperature, R will have the same temperature dependence as ρ. (Examination of the coefficients of linear expansion shows them to be about two orders of magnitude less than typical temperature coefficients of resistivity, and so the effect of temperature on L and A is about two orders of magnitude less than on ρ.) Thus,

Equation:

$$R = R_0(1 + \alpha\Delta T)$$

is the temperature dependence of the resistance of an object, where R_0 is the original resistance and R is the resistance after a temperature change ΔT. Numerous thermometers are based on the effect of temperature on resistance. (See [link].) One of the most common is the thermistor, a semiconductor crystal with a strong temperature dependence, the resistance of which is measured to obtain its temperature. The device is small, so that it quickly comes into thermal equilibrium with the part of a person it touches.



These familiar thermometers are based on the automated measurement of a thermistor’s temperature-dependent resistance. (credit: Biol, Wikimedia Commons)

Example:

Calculating Resistance: Hot-Filament Resistance

Although caution must be used in applying $\rho = \rho_0(1 + \alpha\Delta T)$ and $R = R_0(1 + \alpha\Delta T)$ for temperature changes greater than 100ºC, for tungsten the equations work reasonably well for very large temperature changes. What, then, is the resistance of the tungsten filament in the previous example if its temperature is increased from room temperature (20ºC) to a typical operating temperature of 2850ºC?

Strategy

This is a straightforward application of $R = R_0(1 + \alpha\Delta T)$, since the original resistance of the filament was given to be $R_0 = 0.350\ \Omega$, and the temperature change is $\Delta T = 2830^\circ\text{C}$.

Solution

The hot resistance R is obtained by entering known values into the above equation:

Equation:

$$\begin{aligned} R &= R_0(1 + \alpha\Delta T) \\ &= (0.350\ \Omega)[1 + (4.5\times10^{-3}/^\circ\text{C})(2830^\circ\text{C})] \\ &= 4.8\ \Omega. \end{aligned}$$

Discussion

This value is consistent with the headlight resistance example in Ohm's Law: Resistance and Simple Circuits.

Note:

PhET Explorations: Resistance in a Wire

Learn about the physics of resistance in a wire. Change its resistivity, length, and area to see how they affect the wire's resistance. The sizes of the symbols in the equation change along with the diagram of a wire.

https://phet.colorado.edu/sims/html/resistance-in-a-wire/latest/resistance-in-a-wire_en.html

Section Summary

- The resistance R of a cylinder of length L and cross-sectional area A is $R = \frac{\rho L}{A}$, where ρ is the resistivity of the material.
- Values of ρ in [link] show that materials fall into three groups—*conductors, semiconductors, and insulators*.
- Temperature affects resistivity; for relatively small temperature changes ΔT, resistivity is $\rho = \rho_0(1 + \alpha\Delta T)$, where ρ_0 is the original resistivity and α is the temperature coefficient of resistivity.
- [link] gives values for α, the temperature coefficient of resistivity.
- The resistance R of an object also varies with temperature: $R = R_0(1 + \alpha\Delta T)$, where R_0 is the original resistance, and R is the resistance after the temperature change.

Conceptual Questions

Exercise:

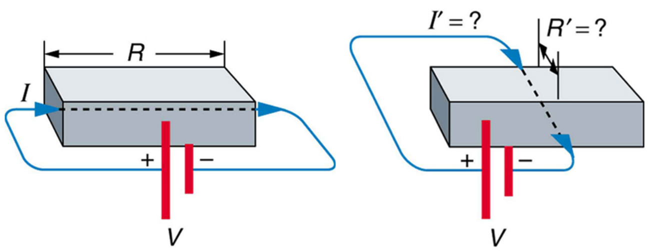
Problem:

In which of the three semiconducting materials listed in [link] do impurities supply free charges? (Hint: Examine the range of resistivity for each and determine whether the pure semiconductor has the higher or lower conductivity.)

Exercise:

Problem:

Does the resistance of an object depend on the path current takes through it? Consider, for example, a rectangular bar—is its resistance the same along its length as across its width? (See [link].)



Does current taking two different paths through the same object encounter different resistance?

Exercise:

Problem:

If aluminum and copper wires of the same length have the same resistance, which has the larger diameter? Why?

Exercise:

Problem:

Explain why $R = R_0(1 + \alpha\Delta T)$ for the temperature variation of the resistance R of an object is not as accurate as $\rho = \rho_0(1 + \alpha\Delta T)$, which gives the temperature variation of resistivity ρ.

Problems & Exercises

Exercise:

Problem:

What is the resistance of a 20.0-m-long piece of 12-gauge copper wire having a 2.053-mm diameter?

Solution:

0.104 Ω

Exercise:

Problem:

The diameter of 0-gauge copper wire is 8.252 mm. Find the resistance of a 1.00-km length of such wire used for power transmission.

Exercise:

Problem:

If the 0.100-mm diameter tungsten filament in a light bulb is to have a resistance of $0.200\ \Omega$ at 20.0°C, how long should it be?

Solution:

$2.8\times10^{-2}\ \text{m}$

Exercise:

Problem:

Find the ratio of the diameter of aluminum to copper wire, if they have the same resistance per unit length (as they might in household wiring).

Exercise:

Problem:

What current flows through a 2.54-cm-diameter rod of pure silicon that is 20.0 cm long, when 1.00×10^{3} V is applied to it? (Such a rod may be used to make nuclear-particle detectors, for example.)

Solution:

$1.10\times10^{-3}\ \text{A}$

Exercise:

Problem:

(a) To what temperature must you raise a copper wire, originally at 20.0°C, to double its resistance, neglecting any changes in dimensions? (b) Does this happen in household wiring under ordinary circumstances?

Exercise:

Problem:

A resistor made of Nichrome wire is used in an application where its resistance cannot change more than 1.00% from its value at 20.0°C. Over what temperature range can it be used?

Solution:

-5°C to 45°C

Exercise:

Problem:

Of what material is a resistor made if its resistance is 40.0% greater at 100°C than at 20.0°C?

Exercise:

Problem:

An electronic device designed to operate at any temperature in the range from $–10.0^{\circ}\text{C}$ to 55.0°C contains pure carbon resistors. By what factor does their resistance increase over this range?

Solution:

1.03

Exercise:

Problem:

(a) Of what material is a wire made, if it is 25.0 m long with a 0.100 mm diameter and has a resistance of $77.7\ \Omega$ at $20.0^{\circ}\mathrm{C}$? (b) What is its resistance at $150^{\circ}\mathrm{C}$?

Exercise:

Problem:

Assuming a constant temperature coefficient of resistivity, what is the maximum percent decrease in the resistance of a constantan wire starting at $20.0^{\circ}\mathrm{C}$?

Solution:

0.06%

Exercise:

Problem:

A wire is drawn through a die, stretching it to four times its original length. By what factor does its resistance increase?

Exercise:

Problem:

A copper wire has a resistance of $0.500\ \Omega$ at $20.0^{\circ}\mathrm{C}$, and an iron wire has a resistance of $0.525\ \Omega$ at the same temperature. At what temperature are their resistances equal?

Solution:

$-17^{\circ}\mathrm{C}$

Exercise:

Problem:

(a) Digital medical thermometers determine temperature by measuring the resistance of a semiconductor device called a thermistor (which has $\alpha = -0.0600/^{\circ}\text{C}$) when it is at the same temperature as the patient. What is a patient's temperature if the thermistor's resistance at that temperature is 82.0% of its value at 37.0°C (normal body temperature)? (b) The negative value for α may not be maintained for very low temperatures. Discuss why and whether this is the case here. (Hint: Resistance can't become negative.)

Exercise:

Problem: Integrated Concepts

(a) Redo [link] taking into account the thermal expansion of the tungsten filament. You may assume a thermal expansion coefficient of $12 \times 10^{-6}/^{\circ}\text{C}$. (b) By what percentage does your answer differ from that in the example?

Solution:

(a) $4.7\ \Omega$ (total)

(b) 3.0% decrease

Exercise:

Problem: Unreasonable Results

(a) To what temperature must you raise a resistor made of constantan to double its resistance, assuming a constant temperature coefficient of resistivity? (b) To cut it in half? (c) What is unreasonable about these results? (d) Which assumptions are unreasonable, or which premises are inconsistent?

Glossary

resistivity
: an intrinsic property of a material, independent of its shape or size, directly proportional to the resistance, denoted by ρ

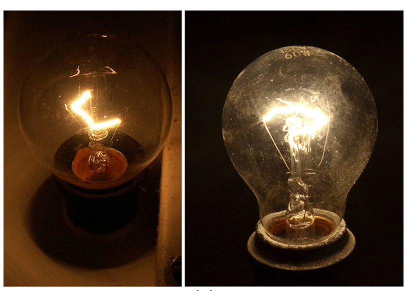
temperature coefficient of resistivity
: an empirical quantity, denoted by α, which describes the change in resistance or resistivity of a material with temperature

Electric Power and Energy

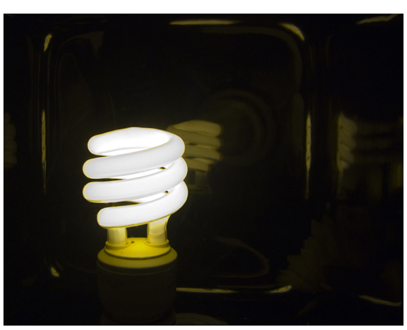
- Calculate the power dissipated by a resistor and power supplied by a power supply.
- Calculate the cost of electricity under various circumstances.

Power in Electric Circuits

Power is associated by many people with electricity. Knowing that power is the rate of energy use or energy conversion, what is the expression for **electric power**? Power transmission lines might come to mind. We also think of lightbulbs in terms of their power ratings in watts. Let us compare a 25-W bulb with a 60-W bulb. (See [link](a).) Since both operate on the same voltage, the 60-W bulb must draw more current to have a greater power rating. Thus the 60-W bulb’s resistance must be lower than that of a 25-W bulb. If we increase voltage, we also increase power. For example, when a 25-W bulb that is designed to operate on 120 V is connected to 240 V, it briefly glows very brightly and then burns out. Precisely how are voltage, current, and resistance related to electric power?



(a)



(b)

(a) Which of these lightbulbs, the 25-W bulb (upper left) or the 60-W bulb (upper right), has the higher resistance? Which draws more current? Which uses the most energy? Can you tell from the color that the 25-W filament is cooler? Is the brighter bulb a different color and if so why? (credits: Dickbauch, Wikimedia Commons; Greg Westfall, Flickr) (b) This compact fluorescent light (CFL) puts out the same intensity of light as the 60-W bulb, but at 1/4 to 1/10 the input power. (credit: dbgg1979, Flickr)

Electric energy depends on both the voltage involved and the charge moved. This is expressed most simply as $\mathrm{PE} = \mathrm{qV}$, where q is the charge moved and V is the voltage (or more precisely, the potential difference the

charge moves through). Power is the rate at which energy is moved, and so electric power is

Equation:

$$P = \frac{\text{PE}}{t} = \frac{\text{qV}}{t}.$$

Recognizing that current is $I = q/t$ (note that $\Delta t = t$ here), the expression for power becomes

Equation:

$$P = \text{IV}.$$

Electric power (P) is simply the product of current times voltage. Power has familiar units of watts. Since the SI unit for potential energy (PE) is the joule, power has units of joules per second, or watts. Thus, $1 \text{ A} \cdot \text{V} = 1 \text{ W}$. For example, cars often have one or more auxiliary power outlets with which you can charge a cell phone or other electronic devices. These outlets may be rated at 20 A, so that the circuit can deliver a maximum power $P = \text{IV} = (20 \text{ A})(12 \text{ V}) = 240 \text{ W}$. In some applications, electric power may be expressed as volt-amperes or even kilovolt-amperes ($1 \text{ kA} \cdot \text{V} = 1 \text{ kW}$).

To see the relationship of power to resistance, we combine Ohm's law with $P = \text{IV}$. Substituting $I = \text{V/R}$ gives $P = (V/R)V = V^2/R$. Similarly, substituting $V = \text{IR}$ gives $P = I(\text{IR}) = I^2R$. Three expressions for electric power are listed together here for convenience:

Equation:

$$P = \text{IV}$$

Equation:

$$P = \frac{V^2}{R}$$

Equation:

$$P = I^2 R.$$

Note that the first equation is always valid, whereas the other two can be used only for resistors. In a simple circuit, with one voltage source and a single resistor, the power supplied by the voltage source and that dissipated by the resistor are identical. (In more complicated circuits, P can be the power dissipated by a single device and not the total power in the circuit.)

Different insights can be gained from the three different expressions for electric power. For example, $P = V^2/R$ implies that the lower the resistance connected to a given voltage source, the greater the power delivered. Furthermore, since voltage is squared in $P = V^2/R$, the effect of applying a higher voltage is perhaps greater than expected. Thus, when the voltage is doubled to a 25-W bulb, its power nearly quadruples to about 100 W, burning it out. If the bulb's resistance remained constant, its power would be exactly 100 W, but at the higher temperature its resistance is higher, too.

Example:
Calculating Power Dissipation and Current: Hot and Cold Power
(a) Consider the examples given in Ohm's Law: Resistance and Simple Circuits and Resistance and Resistivity. Then find the power dissipated by the car headlight in these examples, both when it is hot and when it is cold. (b) What current does it draw when cold?
Strategy for (a)
For the hot headlight, we know voltage and current, so we can use $P = \text{IV}$ to find the power. For the cold headlight, we know the voltage and resistance, so we can use $P = V^2/R$ to find the power.
Solution for (a)
Entering the known values of current and voltage for the hot headlight, we obtain
Equation:

$$P = \text{IV} = (2.50\ \text{A})(12.0\ \text{V}) = 30.0\ \text{W}.$$

The cold resistance was $0.350\ \Omega$, and so the power it uses when first switched on is
Equation:

$$P = \frac{V^2}{R} = \frac{(12.0\ \text{V})^2}{0.350\ \Omega} = 411\ \text{W}.$$

Discussion for (a)
The 30 W dissipated by the hot headlight is typical. But the 411 W when cold is surprisingly higher. The initial power quickly decreases as the bulb's temperature increases and its resistance increases.
Strategy and Solution for (b)
The current when the bulb is cold can be found several different ways. We rearrange one of the power equations, $P = I^2R$, and enter known values, obtaining
Equation:

$$I = \sqrt{\frac{P}{R}} = \sqrt{\frac{411\ \text{W}}{0.350\ \Omega}} = 34.3\ \text{A}.$$

Discussion for (b)
The cold current is remarkably higher than the steady-state value of 2.50 A, but the current will quickly decline to that value as the bulb's temperature increases. Most fuses and circuit breakers (used to limit the current in a circuit) are designed to tolerate very high currents briefly as a device comes on. In some cases, such as with electric motors, the current remains high for several seconds, necessitating special "slow blow" fuses.

The Cost of Electricity

The more electric appliances you use and the longer they are left on, the higher your electric bill. This familiar fact is based on the relationship between energy and power. You pay for the energy used. Since $P = E/t$, we see that
Equation:

$$E = \mathrm{Pt}$$

is the energy used by a device using power P for a time interval t. For example, the more lightbulbs burning, the greater P used; the longer they are on, the greater t is. The energy unit on electric bills is the kilowatt-hour $(\mathrm{kW} \cdot \mathrm{h})$, consistent with the relationship $E = \mathrm{Pt}$. It is easy to estimate the cost of operating electric appliances if you have some idea of their power consumption rate in watts or kilowatts, the time they are on in hours, and the cost per kilowatt-hour for your electric utility. Kilowatt-hours, like all other specialized energy units such as food calories, can be converted to joules. You can prove to yourself that $1\ \mathrm{kW} \cdot \mathrm{h} = 3.6 \times 10^6\ \mathrm{J}$.

The electrical energy (E) used can be reduced either by reducing the time of use or by reducing the power consumption of that appliance or fixture. This will not only reduce the cost, but it will also result in a reduced impact on the environment. Improvements to lighting are some of the fastest ways to reduce the electrical energy used in a home or business. About 20% of a home's use of energy goes to lighting, while the number for commercial establishments is closer to 40%. Fluorescent lights are about four times more efficient than incandescent lights—this is true for both the long tubes and the compact fluorescent lights (CFL). (See [link](b).) Thus, a 60-W incandescent bulb can be replaced by a 15-W CFL, which has the same brightness and color. CFLs have a bent tube inside a globe or a spiral-shaped tube, all connected to a standard screw-in base that fits standard incandescent light sockets. (Original problems with color, flicker, shape, and high initial investment for CFLs have been addressed in recent years.) The heat transfer from these CFLs is less, and they last up to 10 times longer. The significance of an investment in such bulbs is addressed in the next example. New white LED lights (which are clusters of small LED bulbs) are even more efficient (twice that of CFLs) and last 5 times longer than CFLs. However, their cost is still high.

Note:
Making Connections: Energy, Power, and Time

The relationship $E = \text{Pt}$ is one that you will find useful in many different contexts. The energy your body uses in exercise is related to the power level and duration of your activity, for example. The amount of heating by a power source is related to the power level and time it is applied. Even the radiation dose of an X-ray image is related to the power and time of exposure.

Example:

Calculating the Cost Effectiveness of Compact Fluorescent Lights (CFL)

If the cost of electricity in your area is 12 cents per kWh, what is the total cost (capital plus operation) of using a 60-W incandescent bulb for 1000 hours (the lifetime of that bulb) if the bulb cost 25 cents? (b) If we replace this bulb with a compact fluorescent light that provides the same light output, but at one-quarter the wattage, and which costs \$1.50 but lasts 10 times longer (10,000 hours), what will that total cost be?

Strategy

To find the operating cost, we first find the energy used in kilowatt-hours and then multiply by the cost per kilowatt-hour.

Solution for (a)

The energy used in kilowatt-hours is found by entering the power and time into the expression for energy:

Equation:

$$E = \text{Pt} = (60\text{ W})(1000\text{ h}) = 60{,}000\text{ W}\cdot\text{h}.$$

In kilowatt-hours, this is

Equation:

$$E = 60.0\text{ kW}\cdot\text{h}.$$

Now the electricity cost is

Equation:

$$\text{cost} = (60.0\text{ kW}\cdot\text{h})(\$0.12/\text{kW}\cdot\text{h}) = \$7.20.$$

The total cost will be \$7.20 for 1000 hours (about one-half year at 5 hours per day).

Solution for (b)

Since the CFL uses only 15 W and not 60 W, the electricity cost will be \$7.20/4 = \$1.80. The CFL will last 10 times longer than the incandescent, so that the investment cost will be 1/10 of the bulb cost for that time period of use, or 0.1(\$1.50) = \$0.15. Therefore, the total cost will be \$1.95 for 1000 hours.

Discussion

Therefore, it is much cheaper to use the CFLs, even though the initial investment is higher. The increased cost of labor that a business must include for replacing the incandescent bulbs more often has not been figured in here.

Note:

Making Connections: Take-Home Experiment—Electrical Energy Use Inventory

1) Make a list of the power ratings on a range of appliances in your home or room. Explain why something like a toaster has a higher rating than a digital clock. Estimate the energy consumed by these appliances in an average day (by estimating their time of use). Some appliances might only state the operating current. If the household voltage is 120 V, then use $P = \text{IV}$. 2) Check out the total wattage used in the rest rooms of your school's floor or building. (You might need to assume the long fluorescent lights in use are rated at 32 W.) Suppose that the building was closed all weekend and that these lights were left on from 6 p.m. Friday until 8 a.m. Monday. What would this oversight cost? How about for an entire year of weekends?

Section Summary

- Electric power P is the rate (in watts) that energy is supplied by a source or dissipated by a device.

- Three expressions for electrical power are
  Equation:

  $$P = \text{IV},$$

  Equation:

  $$P = \frac{V^2}{R},$$

  and
  Equation:

  $$P = I^2 R.$$

- The energy used by a device with a power P over a time t is $E = \text{Pt}$.

Conceptual Questions

Exercise:

Problem:

Why do incandescent lightbulbs grow dim late in their lives, particularly just before their filaments break?

Exercise:

Problem:

The power dissipated in a resistor is given by $P = V^2/R$, which means power decreases if resistance increases. Yet this power is also given by $P = I^2 R$, which means power increases if resistance increases. Explain why there is no contradiction here.

Problem Exercises

Exercise:

Problem:

What is the power of a 1.00×10^{2} MV lightning bolt having a current of 2.00×10^{4} A?

Solution:

$2.00{\times}10^{12}$ W

Exercise:

Problem:

What power is supplied to the starter motor of a large truck that draws 250 A of current from a 24.0-V battery hookup?

Exercise:

Problem:

A charge of 4.00 C of charge passes through a pocket calculator’s solar cells in 4.00 h. What is the power output, given the calculator’s voltage output is 3.00 V? (See [link].)

The strip of solar cells just above the keys of this calculator convert

light to electricity to supply its energy needs. (credit: Evan-Amos, Wikimedia Commons)

Exercise:

Problem:

How many watts does a flashlight that has 6.00×10^2 C pass through it in 0.500 h use if its voltage is 3.00 V?

Exercise:

Problem:

Find the power dissipated in each of these extension cords: (a) an extension cord having a $0.0600 - \Omega$ resistance and through which 5.00 A is flowing; (b) a cheaper cord utilizing thinner wire and with a resistance of $0.300\ \Omega$.

Solution:

(a) 1.50 W

(b) 7.50 W

Exercise:

Problem:

Verify that the units of a volt-ampere are watts, as implied by the equation $P = \text{IV}$.

Exercise:

Problem:

Show that the units $1\ \mathrm{V}^2/\Omega = 1\mathrm{W}$, as implied by the equation $P = V^2/R$.

Solution:

$$\frac{V^2}{\Omega} = \frac{V^2}{\mathrm{V/A}} = \mathrm{AV} = \left(\frac{C}{s}\right)\left(\frac{J}{C}\right) = \frac{J}{s} = 1\ \mathrm{W}$$

Exercise:

Problem:

Show that the units $1\ \mathrm{A}^2 \cdot \Omega = 1\ \mathrm{W}$, as implied by the equation $P = I^2R$.

Exercise:

Problem:

Verify the energy unit equivalence that $1\ \mathrm{kW} \cdot \mathrm{h} = 3.60{\times}10^6\ \mathrm{J}$.

Solution:

$$1\ \mathrm{kW} \cdot \mathrm{h}{=} \left(\frac{1{\times}10^3\ \mathrm{J}}{1\ \mathrm{s}}\right)(1\ \mathrm{h})\left(\frac{3600\ \mathrm{s}}{1\ \mathrm{h}}\right) = 3.60 \times 10^6\ \mathrm{J}$$

Exercise:

Problem:

Electrons in an X-ray tube are accelerated through $1.00 \times 10^2\ \mathrm{kV}$ and directed toward a target to produce X-rays. Calculate the power of the electron beam in this tube if it has a current of 15.0 mA.

Exercise:

Problem:

An electric water heater consumes 5.00 kW for 2.00 h per day. What is the cost of running it for one year if electricity costs $12.0\ \text{cents/kW}\cdot\text{h}$? See [link].

On-demand electric hot water heater. Heat is supplied to water only when needed. (credit: aviddavid, Flickr)

Solution:

$438/y

Exercise:

Problem:

With a 1200-W toaster, how much electrical energy is needed to make a slice of toast (cooking time = 1 minute)? At $9.0\ \text{cents/kW}\cdot\text{h}$, how much does this cost?

Exercise:

Problem:

What would be the maximum cost of a CFL such that the total cost (investment plus operating) would be the same for both CFL and incandescent 60-W bulbs? Assume the cost of the incandescent bulb is 25 cents and that electricity costs $10\ \text{cents/kWh}$. Calculate the cost for 1000 hours, as in the cost effectiveness of CFL example.

Solution:

$6.25

Exercise:

Problem:

Some makes of older cars have 6.00-V electrical systems. (a) What is the hot resistance of a 30.0-W headlight in such a car? (b) What current flows through it?

Exercise:

Problem:

Alkaline batteries have the advantage of putting out constant voltage until very nearly the end of their life. How long will an alkaline battery rated at $1.00\ \text{A}\cdot\text{h}$ and 1.58 V keep a 1.00-W flashlight bulb burning?

Solution:

1.58 h

Exercise:

Problem:

A cauterizer, used to stop bleeding in surgery, puts out 2.00 mA at 15.0 kV. (a) What is its power output? (b) What is the resistance of the path?

Exercise:

Problem:

The average television is said to be on 6 hours per day. Estimate the yearly cost of electricity to operate 100 million TVs, assuming their power consumption averages 150 W and the cost of electricity averages $12.0 \text{ cents/kW} \cdot \text{h}$.

Solution:

$3.94 billion/year

Exercise:

Problem:

An old lightbulb draws only 50.0 W, rather than its original 60.0 W, due to evaporative thinning of its filament. By what factor is its diameter reduced, assuming uniform thinning along its length? Neglect any effects caused by temperature differences.

Exercise:

Problem:

00-gauge copper wire has a diameter of 9.266 mm. Calculate the power loss in a kilometer of such wire when it carries $1.00 \times 10^{2} \text{ A}$.

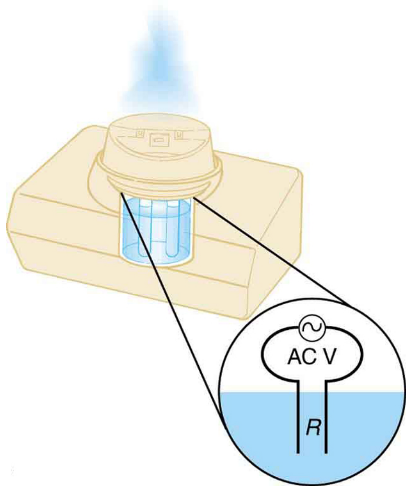
Solution:

25.5 W

Exercise:

Problem: **Integrated Concepts**

Cold vaporizers pass a current through water, evaporating it with only a small increase in temperature. One such home device is rated at 3.50 A and utilizes 120 V AC with 95.0% efficiency. (a) What is the vaporization rate in grams per minute? (b) How much water must you put into the vaporizer for 8.00 h of overnight operation? (See [link].)



This cold vaporizer passes current directly through water, vaporizing it directly with relatively little temperature increase.

Exercise:

Problem: Integrated Concepts

(a) What energy is dissipated by a lightning bolt having a 20,000-A current, a voltage of 1.00×10^2 MV, and a length of 1.00 ms? (b) What mass of tree sap could be raised from $18.0^{\circ}C$ to its boiling point and then evaporated by this energy, assuming sap has the same thermal characteristics as water?

Solution:

(a) 2.00×10^9 J

(b) 769 kg

Exercise:

Problem: Integrated Concepts

What current must be produced by a 12.0-V battery-operated bottle warmer in order to heat 75.0 g of glass, 250 g of baby formula, and 3.00×10^2 g of aluminum from $20.0^{\circ}C$ to $90.0^{\circ}C$ in 5.00 min?

Exercise:

Problem: Integrated Concepts

How much time is needed for a surgical cauterizer to raise the temperature of 1.00 g of tissue from $37.0^{\circ}C$ to $100^{\circ}C$ and then boil away 0.500 g of water, if it puts out 2.00 mA at 15.0 kV? Ignore heat transfer to the surroundings.

Solution:

45.0 s

Exercise:

Problem: Integrated Concepts

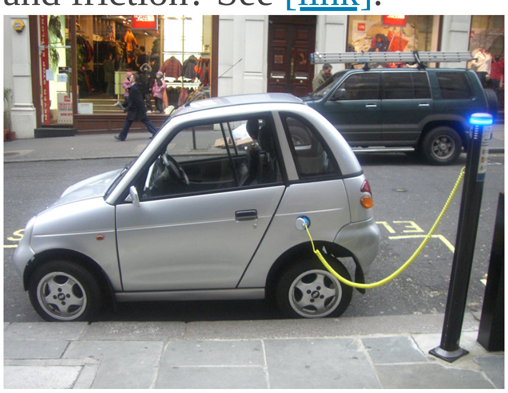
Hydroelectric generators (see [link]) at Hoover Dam produce a maximum current of 8.00×10^3 A at 250 kV. (a) What is the power output? (b) The water that powers the generators enters and leaves the system at low speed (thus its kinetic energy does not change) but loses 160 m in altitude. How many cubic meters per second are needed, assuming 85.0% efficiency?

Hydroelectric generators
at the Hoover dam.
(credit: Jon Sullivan)

Exercise:

Problem: Integrated Concepts

(a) Assuming 95.0% efficiency for the conversion of electrical power by the motor, what current must the 12.0-V batteries of a 750-kg electric car be able to supply: (a) To accelerate from rest to 25.0 m/s in 1.00 min? (b) To climb a $2.00 \times 10^2\text{-m}$-high hill in 2.00 min at a constant 25.0-m/s speed while exerting $5.00 \times 10^2\ \text{N}$ of force to overcome air resistance and friction? (c) To travel at a constant 25.0-m/s speed, exerting a $5.00 \times 10^2\ \text{N}$ force to overcome air resistance and friction? See [link].



This REVAi, an electric

car, gets recharged on a street in London. (credit: Frank Hebbert)

Solution:

(a) 343 A

(b) 2.17×10^3 A

(c) 1.10×10^3 A

Exercise:

Problem: Integrated Concepts

A light-rail commuter train draws 630 A of 650-V DC electricity when accelerating. (a) What is its power consumption rate in kilowatts? (b) How long does it take to reach 20.0 m/s starting from rest if its loaded mass is 5.30×10^4 kg, assuming 95.0% efficiency and constant power? (c) Find its average acceleration. (d) Discuss how the acceleration you found for the light-rail train compares to what might be typical for an automobile.

Exercise:

Problem: Integrated Concepts

(a) An aluminum power transmission line has a resistance of $0.0580\ \Omega/\text{km}$. What is its mass per kilometer? (b) What is the mass per kilometer of a copper line having the same resistance? A lower resistance would shorten the heating time. Discuss the practical limits to speeding the heating by lowering the resistance.

Solution:

(a) 1.23×10^3 kg

(b) 2.64×10^3 kg

Exercise:

Problem: **Integrated Concepts**

(a) An immersion heater utilizing 120 V can raise the temperature of a 1.00×10^2-g aluminum cup containing 350 g of water from $20.0^{\circ}C$ to $95.0^{\circ}C$ in 2.00 min. Find its resistance, assuming it is constant during the process. (b) A lower resistance would shorten the heating time. Discuss the practical limits to speeding the heating by lowering the resistance.

Exercise:

Problem: **Integrated Concepts**

(a) What is the cost of heating a hot tub containing 1500 kg of water from $10.0^{\circ}C$ to $40.0^{\circ}C$, assuming 75.0% efficiency to account for heat transfer to the surroundings? The cost of electricity is $9 \text{ cents/kW} \cdot \text{h}$. (b) What current was used by the 220-V AC electric heater, if this took 4.00 h?

Exercise:

Problem: **Unreasonable Results**

(a) What current is needed to transmit 1.00×10^2 MW of power at 480 V? (b) What power is dissipated by the transmission lines if they have a 1.00 - Ω resistance? (c) What is unreasonable about this result? (d) Which assumptions are unreasonable, or which premises are inconsistent?

Solution:

(a) 2.08×10^5 A

(b) 4.33×10^4 MW

(c) The transmission lines dissipate more power than they are supposed to transmit.

(d) A voltage of 480 V is unreasonably low for a transmission voltage. Long-distance transmission lines are kept at much higher voltages (often hundreds of kilovolts) to reduce power losses.

Exercise:

Problem: Unreasonable Results

(a) What current is needed to transmit 1.00×10^2 MW of power at 10.0 kV? (b) Find the resistance of 1.00 km of wire that would cause a 0.0100% power loss. (c) What is the diameter of a 1.00-km-long copper wire having this resistance? (d) What is unreasonable about these results? (e) Which assumptions are unreasonable, or which premises are inconsistent?

Exercise:

Problem: Construct Your Own Problem

Consider an electric immersion heater used to heat a cup of water to make tea. Construct a problem in which you calculate the needed resistance of the heater so that it increases the temperature of the water and cup in a reasonable amount of time. Also calculate the cost of the electrical energy used in your process. Among the things to be considered are the voltage used, the masses and heat capacities involved, heat losses, and the time over which the heating takes place. Your instructor may wish for you to consider a thermal safety switch (perhaps bimetallic) that will halt the process before damaging temperatures are reached in the immersion unit.

Glossary

electric power

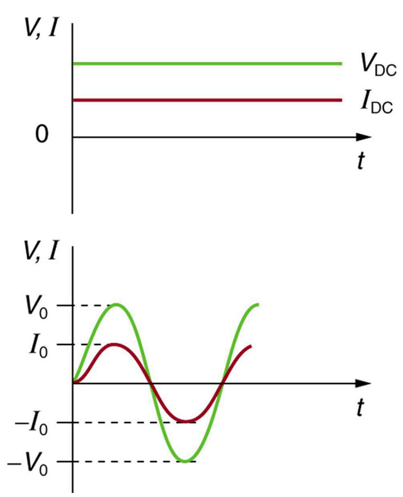
the rate at which electrical energy is supplied by a source or dissipated by a device; it is the product of current times voltage

Alternating Current versus Direct Current

- Explain the differences and similarities between AC and DC current.
- Calculate rms voltage, current, and average power.
- Explain why AC current is used for power transmission.

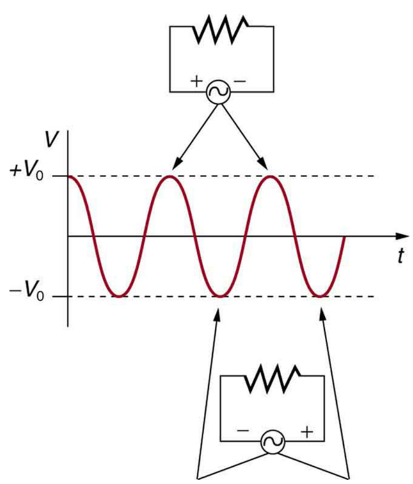
Alternating Current

Most of the examples dealt with so far, and particularly those utilizing batteries, have constant voltage sources. Once the current is established, it is thus also a constant. **Direct current** (DC) is the flow of electric charge in only one direction. It is the steady state of a constant-voltage circuit. Most well-known applications, however, use a time-varying voltage source. **Alternating current** (AC) is the flow of electric charge that periodically reverses direction. If the source varies periodically, particularly sinusoidally, the circuit is known as an alternating current circuit. Examples include the commercial and residential power that serves so many of our needs. [link] shows graphs of voltage and current versus time for typical DC and AC power. The AC voltages and frequencies commonly used in homes and businesses vary around the world.



(a) DC voltage and current are constant in time, once the

current is established. (b) A graph of voltage and current versus time for 60-Hz AC power. The voltage and current are sinusoidal and are in phase for a simple resistance circuit. The frequencies and peak voltages of AC sources differ greatly.



The potential difference V between the terminals of an AC voltage source fluctuates as

shown. The mathematical expression for V is given by $V = V_0 \sin 2\pi \mathrm{ft}$.

[link] shows a schematic of a simple circuit with an AC voltage source. The voltage between the terminals fluctuates as shown, with the **AC voltage** given by

Equation:

$$V = V_0 \sin 2\pi \mathrm{ft},$$

where V is the voltage at time t, V_0 is the peak voltage, and f is the frequency in hertz. For this simple resistance circuit, $I = \mathrm{V/R}$, and so the **AC current** is

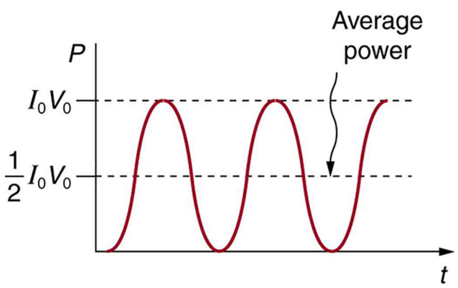
Equation:

$$I = I_0 \sin 2\pi \mathrm{ft},$$

where I is the current at time t, and $I_0 = V_0/\mathrm{R}$ is the peak current. For this example, the voltage and current are said to be in phase, as seen in [link](b).

Current in the resistor alternates back and forth just like the driving voltage, since $I = \mathrm{V/R}$. If the resistor is a fluorescent light bulb, for example, it brightens and dims 120 times per second as the current repeatedly goes through zero. A 120-Hz flicker is too rapid for your eyes to detect, but if you wave your hand back and forth between your face and a fluorescent light, you will see a stroboscopic effect evidencing AC. The fact that the light output fluctuates means that the power is fluctuating. The power supplied is $P = \mathrm{IV}$. Using the expressions for I and V above, we see that the time dependence of power is $P = I_0 V_0 \sin^2 2\pi \mathrm{ft}$, as shown in [link].

Note:
Making Connections: Take-Home Experiment—AC/DC Lights
Wave your hand back and forth between your face and a fluorescent light bulb. Do you observe the same thing with the headlights on your car? Explain what you observe. *Warning: Do not look directly at very bright light*.



AC power as a function of time. Since the voltage and current are in phase here, their product is non-negative and fluctuates between zero and I_0V_0. Average power is $(1/2)I_0V_0$.

We are most often concerned with average power rather than its fluctuations—that 60-W light bulb in your desk lamp has an average power consumption of 60 W, for example. As illustrated in [link], the average power P_{ave} is

Equation:

$$P_{\text{ave}} = \frac{1}{2}I_0V_0.$$

This is evident from the graph, since the areas above and below the $(1/2)I_0V_0$ line are equal, but it can also be proven using trigonometric identities. Similarly, we define an average or **rms current** I_{rms} and average or **rms voltage** V_{rms} to be, respectively,

Equation:

$$I_{\text{rms}} = \frac{I_0}{\sqrt{2}}$$

and

Equation:

$$V_{\text{rms}} = \frac{V_0}{\sqrt{2}}.$$

where rms stands for root mean square, a particular kind of average. In general, to obtain a root mean square, the particular quantity is squared, its mean (or average) is found, and the square root is taken. This is useful for AC, since the average value is zero. Now,

Equation:

$$P_{\text{ave}} = I_{\text{rms}}V_{\text{rms}},$$

which gives

Equation:

$$P_{\text{ave}} = \frac{I_0}{\sqrt{2}} \cdot \frac{V_0}{\sqrt{2}} = \frac{1}{2}I_0V_0,$$

as stated above. It is standard practice to quote I_{rms}, V_{rms}, and P_{ave} rather than the peak values. For example, most household electricity is 120 V AC, which means that V_{rms} is 120 V. The common 10-A circuit breaker will interrupt a sustained I_{rms} greater than 10 A. Your 1.0-kW microwave oven

consumes $P_{\text{ave}} = 1.0\ \text{kW}$, and so on. You can think of these rms and average values as the equivalent DC values for a simple resistive circuit.

To summarize, when dealing with AC, Ohm's law and the equations for power are completely analogous to those for DC, but rms and average values are used for AC. Thus, for AC, Ohm's law is written

Equation:

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{R}.$$

The various expressions for AC power P_{ave} are

Equation:

$$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}},$$

Equation:

$$P_{\text{ave}} = \frac{V_{\text{rms}}^2}{R},$$

and

Equation:

$$P_{\text{ave}} = I_{\text{rms}}^2 R.$$

Example:

Peak Voltage and Power for AC

(a) What is the value of the peak voltage for 120-V AC power? (b) What is the peak power consumption rate of a 60.0-W AC light bulb?

Strategy

We are told that V_{rms} is 120 V and P_{ave} is 60.0 W. We can use $V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$ to find the peak voltage, and we can manipulate the definition of power to

find the peak power from the given average power.
Solution for (a)
Solving the equation $V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$ for the peak voltage V_0 and substituting the known value for V_{rms} gives
Equation:

$$V_0 = \sqrt{2}V_{\text{rms}} = 1.414(120\text{ V}) = 170\text{ V}.$$

Discussion for (a)
This means that the AC voltage swings from 170 V to –170 V and back 60 times every second. An equivalent DC voltage is a constant 120 V.
Solution for (b)
Peak power is peak current times peak voltage. Thus,
Equation:

$$P_0 = I_0V_0 = 2\left(\frac{1}{2}I_0V_0\right) = 2P_{\text{ave}}.$$

We know the average power is 60.0 W, and so
Equation:

$$P_0 = 2(60.0\text{ W}) = 120\text{ W}.$$

Discussion
So the power swings from zero to 120 W one hundred twenty times per second (twice each cycle), and the power averages 60 W.

Why Use AC for Power Distribution?

Most large power-distribution systems are AC. Moreover, the power is transmitted at much higher voltages than the 120-V AC (240 V in most parts of the world) we use in homes and on the job. Economies of scale make it cheaper to build a few very large electric power-generation plants than to build numerous small ones. This necessitates sending power long distances, and it is obviously important that energy losses en route be

minimized. High voltages can be transmitted with much smaller power losses than low voltages, as we shall see. (See [link].) For safety reasons, the voltage at the user is reduced to familiar values. The crucial factor is that it is much easier to increase and decrease AC voltages than DC, so AC is used in most large power distribution systems.

Power is distributed over large distances at high voltage to reduce power loss in the transmission lines. The voltages generated at the power plant are stepped up by passive devices called transformers (see Transformers) to 330,000 volts (or more in some places worldwide). At the point of use, the transformers reduce the voltage transmitted for safe residential and commercial use. (Credit: GeorgHH, Wikimedia Commons)

Example:

Power Losses Are Less for High-Voltage Transmission

(a) What current is needed to transmit 100 MW of power at 200 kV? (b) What is the power dissipated by the transmission lines if they have a resistance of 1.00 Ω? (c) What percentage of the power is lost in the transmission lines?

Strategy

We are given $P_{\text{ave}} = 100\text{ MW}$, $V_{\text{rms}} = 200\text{ kV}$, and the resistance of the lines is $R = 1.00\ \Omega$. Using these givens, we can find the current flowing (from $P = \text{IV}$) and then the power dissipated in the lines ($P = I^2R$), and we take the ratio to the total power transmitted.

Solution

To find the current, we rearrange the relationship $P_{\text{ave}} = I_{\text{rms}}V_{\text{rms}}$ and substitute known values. This gives

Equation:

$$I_{\text{rms}} = \frac{P_{\text{ave}}}{V_{\text{rms}}} = \frac{100 \times 10^6\text{ W}}{200 \times 10^3\text{ V}} = 500\text{ A.}$$

Solution

Knowing the current and given the resistance of the lines, the power dissipated in them is found from $P_{\text{ave}} = I_{\text{rms}}^2 R$. Substituting the known values gives

Equation:

$$P_{\text{ave}} = I_{\text{rms}}^2 R = (500\text{ A})^2 (1.00\ \Omega) = 250\text{ kW.}$$

Solution

The percent loss is the ratio of this lost power to the total or input power, multiplied by 100:

Equation:

$$\% \text{ loss} = \frac{250 \text{ kW}}{100 \text{ MW}} \times 100 = 0.250 \ \%.$$

Discussion

One-fourth of a percent is an acceptable loss. Note that if 100 MW of power had been transmitted at 25 kV, then a current of 4000 A would have been needed. This would result in a power loss in the lines of 16.0 MW, or 16.0% rather than 0.250%. The lower the voltage, the more current is needed, and the greater the power loss in the fixed-resistance transmission lines. Of course, lower-resistance lines can be built, but this requires larger and more expensive wires. If superconducting lines could be economically produced, there would be no loss in the transmission lines at all. But, as we shall see in a later chapter, there is a limit to current in superconductors, too. In short, high voltages are more economical for transmitting power, and AC voltage is much easier to raise and lower, so that AC is used in most large-scale power distribution systems.

It is widely recognized that high voltages pose greater hazards than low voltages. But, in fact, some high voltages, such as those associated with common static electricity, can be harmless. So it is not voltage alone that determines a hazard. It is not so widely recognized that AC shocks are often more harmful than similar DC shocks. Thomas Edison thought that AC shocks were more harmful and set up a DC power-distribution system in New York City in the late 1800s. There were bitter fights, in particular between Edison and George Westinghouse and Nikola Tesla, who were advocating the use of AC in early power-distribution systems. AC has prevailed largely due to transformers and lower power losses with high-voltage transmission.

Note:

PhET Explorations: Generator

Generate electricity with a bar magnet! Discover the physics behind the phenomena by exploring magnets and how you can use them to make a bulb light.

Generator

Section Summary

- Direct current (DC) is the flow of electric current in only one direction. It refers to systems where the source voltage is constant.
- The voltage source of an alternating current (AC) system puts out $V = V_0 \sin 2\pi \text{ft}$, where V is the voltage at time t, V_0 is the peak voltage, and f is the frequency in hertz.
- In a simple circuit, $I = \text{V}/\text{R}$ and AC current is $I = I_0 \sin 2\pi \text{ft}$, where I is the current at time t, and $I_0 = V_0/\text{R}$ is the peak current.
- The average AC power is $P_{\text{ave}} = \frac{1}{2} I_0 V_0$.
- Average (rms) current I_{rms} and average (rms) voltage V_{rms} are $I_{\text{rms}} = \frac{I_0}{\sqrt{2}}$ and $V_{\text{rms}} = \frac{V_0}{\sqrt{2}}$, where rms stands for root mean square.
- Thus, $P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}}$.
- Ohm's law for AC is $I_{\text{rms}} = \frac{V_{\text{rms}}}{R}$.
- Expressions for the average power of an AC circuit are $P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}}$, $P_{\text{ave}} = \frac{V_{\text{rms}}^{2}}{R}$, and $P_{\text{ave}} = I_{\text{rms}}^{2} R$, analogous to the expressions for DC circuits.

Conceptual Questions

Exercise:

Problem:

Give an example of a use of AC power other than in the household. Similarly, give an example of a use of DC power other than that supplied by batteries.

Exercise:

Problem:

Why do voltage, current, and power go through zero 120 times per second for 60-Hz AC electricity?

Exercise:

Problem:

You are riding in a train, gazing into the distance through its window. As close objects streak by, you notice that the nearby fluorescent lights make *dashed* streaks. Explain.

Problem Exercises

Exercise:

Problem:

(a) What is the hot resistance of a 25-W light bulb that runs on 120-V AC? (b) If the bulb's operating temperature is $2700^\circ C$, what is its resistance at $2600^\circ C$?

Exercise:

Problem:

Certain heavy industrial equipment uses AC power that has a peak voltage of 679 V. What is the rms voltage?

Solution:

480 V

Exercise:

Problem:

A certain circuit breaker trips when the rms current is 15.0 A. What is the corresponding peak current?

Exercise:

Problem:

Military aircraft use 400-Hz AC power, because it is possible to design lighter-weight equipment at this higher frequency. What is the time for one complete cycle of this power?

Solution:

2.50 ms

Exercise:

Problem:

A North American tourist takes his 25.0-W, 120-V AC razor to Europe, finds a special adapter, and plugs it into 240 V AC. Assuming constant resistance, what power does the razor consume as it is ruined?

Exercise:

Problem:

In this problem, you will verify statements made at the end of the power losses for [link]. (a) What current is needed to transmit 100 MW of power at a voltage of 25.0 kV? (b) Find the power loss in a 1.00 - Ω transmission line. (c) What percent loss does this represent?

Solution:

(a) 4.00 kA

(b) 16.0 MW

(c) 16.0%

Exercise:

Problem:

A small office-building air conditioner operates on 408-V AC and consumes 50.0 kW. (a) What is its effective resistance? (b) What is the cost of running the air conditioner during a hot summer month when it is on 8.00 h per day for 30 days and electricity costs $9.00\ \text{cents/kW}\cdot\text{h}$?

Exercise:

Problem:

What is the peak power consumption of a 120-V AC microwave oven that draws 10.0 A?

Solution:

2.40 kW

Exercise:

Problem:

What is the peak current through a 500-W room heater that operates on 120-V AC power?

Exercise:

Problem:

Two different electrical devices have the same power consumption, but one is meant to be operated on 120-V AC and the other on 240-V AC. (a) What is the ratio of their resistances? (b) What is the ratio of their currents? (c) Assuming its resistance is unaffected, by what factor will the power increase if a 120-V AC device is connected to 240-V AC?

Solution:

(a) 4.0

(b) 0.50

(c) 4.0

Exercise:

Problem:

Nichrome wire is used in some radiative heaters. (a) Find the resistance needed if the average power output is to be 1.00 kW utilizing 120-V AC. (b) What length of Nichrome wire, having a cross-sectional area of 5.00mm^2, is needed if the operating temperature is 500^o C? (c) What power will it draw when first switched on?

Exercise:

Problem:

Find the time after $t = 0$ when the instantaneous voltage of 60-Hz AC first reaches the following values: (a) $V_0/2$ (b) V_0 (c) 0.

Solution:

(a) 1.39 ms

(b) 4.17 ms

(c) 8.33 ms

Exercise:

Problem:

(a) At what two times in the first period following $t = 0$ does the instantaneous voltage in 60-Hz AC equal V_rms? (b) $-V_\text{rms}$?

Glossary

direct current
: (DC) the flow of electric charge in only one direction

alternating current
: (AC) the flow of electric charge that periodically reverses direction

AC voltage

voltage that fluctuates sinusoidally with time, expressed as $V = V_0 \sin 2\pi ft$, where V is the voltage at time t, V_0 is the peak voltage, and f is the frequency in hertz

AC current

current that fluctuates sinusoidally with time, expressed as $I = I_0 \sin 2\pi ft$, where I is the current at time t, I_0 is the peak current, and f is the frequency in hertz

rms current

the root mean square of the current, $I_{\text{rms}} = I_0/\sqrt{2}$, where I_0 is the peak current, in an AC system

rms voltage

the root mean square of the voltage, $V_{\text{rms}} = V_0/\sqrt{2}$, where V_0 is the peak voltage, in an AC system

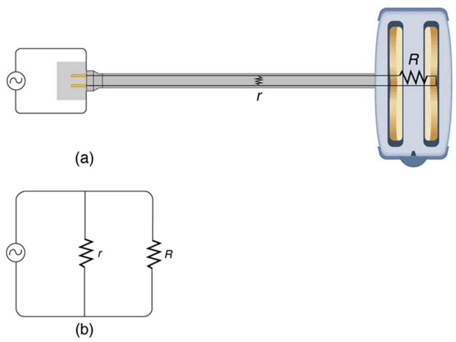
Electric Hazards and the Human Body

- Define thermal hazard, shock hazard, and short circuit.
- Explain what effects various levels of current have on the human body.

There are two known hazards of electricity—thermal and shock. A **thermal hazard** is one where excessive electric power causes undesired thermal effects, such as starting a fire in the wall of a house. A **shock hazard** occurs when electric current passes through a person. Shocks range in severity from painful, but otherwise harmless, to heart-stopping lethality. This section considers these hazards and the various factors affecting them in a quantitative manner. Electrical Safety: Systems and Devices will consider systems and devices for preventing electrical hazards.

Thermal Hazards

Electric power causes undesired heating effects whenever electric energy is converted to thermal energy at a rate faster than it can be safely dissipated. A classic example of this is the **short circuit**, a low-resistance path between terminals of a voltage source. An example of a short circuit is shown in [link]. Insulation on wires leading to an appliance has worn through, allowing the two wires to come into contact. Such an undesired contact with a high voltage is called a *short*. Since the resistance of the short, r, is very small, the power dissipated in the short, $P = V^2/r$, is very large. For example, if V is 120 V and r is $0.100\ \Omega$, then the power is 144 kW, *much* greater than that used by a typical household appliance. Thermal energy delivered at this rate will very quickly raise the temperature of surrounding materials, melting or perhaps igniting them.

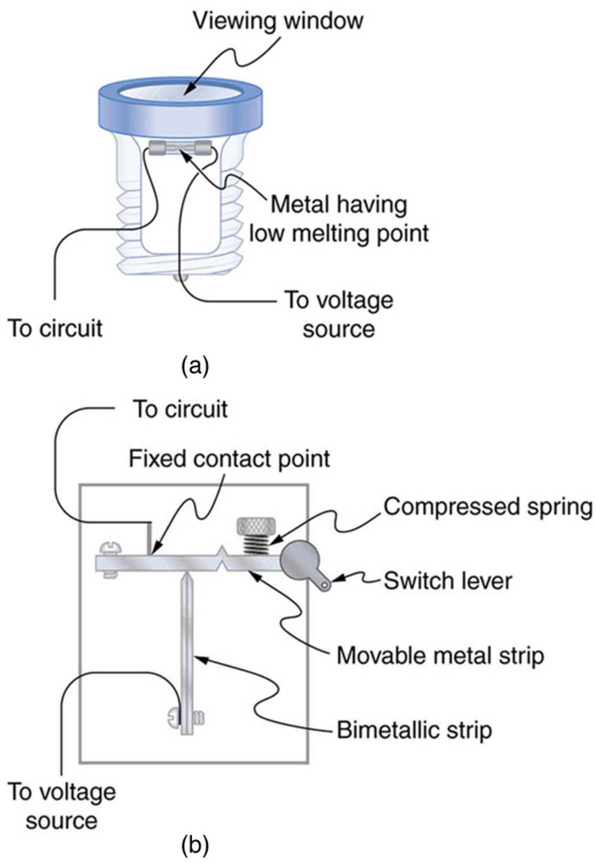


A short circuit is an undesired low-resistance path across a voltage source. (a) Worn insulation on the wires of a toaster allow them to come into contact with a low resistance r. Since $P = V^2/r$, thermal power is created so rapidly that the cord melts or burns. (b) A schematic of the short circuit.

One particularly insidious aspect of a short circuit is that its resistance may actually be decreased due to the increase in temperature. This can happen if the short creates ionization. These charged atoms and molecules are free to move and, thus, lower the resistance r. Since $P = V^2/r$, the power dissipated in the short rises, possibly causing more ionization, more power, and so on. High voltages, such as the 480-V AC used in some industrial applications, lend themselves to this hazard, because higher voltages create higher initial power production in a short.

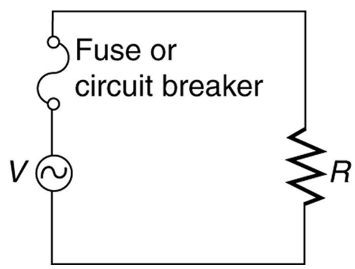
Another serious, but less dramatic, thermal hazard occurs when wires supplying power to a user are overloaded with too great a current. As

discussed in the previous section, the power dissipated in the supply wires is $P = I^2 R_w$, where R_w is the resistance of the wires and I the current flowing through them. If either I or R_w is too large, the wires overheat. For example, a worn appliance cord (with some of its braided wires broken) may have $R_w = 2.00 \ \Omega$ rather than the $0.100 \ \Omega$ it should be. If 10.0 A of current passes through the cord, then $P = I^2 R_w = 200 \text{ W}$ is dissipated in the cord—much more than is safe. Similarly, if a wire with a $0.100 \text{-} \Omega$ resistance is meant to carry a few amps, but is instead carrying 100 A, it will severely overheat. The power dissipated in the wire will in that case be $P = 1000 \text{ W}$. Fuses and circuit breakers are used to limit excessive currents. (See [link] and [link].) Each device opens the circuit automatically when a sustained current exceeds safe limits.



(a) A fuse has a metal strip with a low melting point that, when overheated by an excessive

current, permanently breaks the connection of a circuit to a voltage source. (b) A circuit breaker is an automatic but restorable electric switch. The one shown here has a bimetallic strip that bends to the right and into the notch if overheated. The spring then forces the metal strip downward, breaking the electrical connection at the points.



Schematic of a circuit with a fuse or circuit breaker in it. Fuses and circuit breakers act like automatic switches that open when sustained current exceeds desired limits.

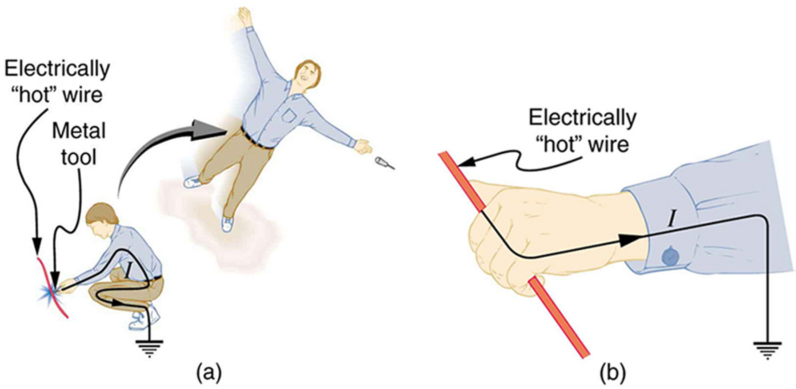
Fuses and circuit breakers for typical household voltages and currents are relatively simple to produce, but those for large voltages and currents experience special problems. For example, when a circuit breaker tries to interrupt the flow of high-voltage electricity, a spark can jump across its points that ionizes the air in the gap and allows the current to continue flowing. Large circuit breakers found in power-distribution systems employ insulating gas and even use jets of gas to blow out such sparks. Here AC is safer than DC, since AC current goes through zero 120 times per second, giving a quick opportunity to extinguish these arcs.

Shock Hazards

Electrical currents through people produce tremendously varied effects. An electrical current can be used to block back pain. The possibility of using electrical current to stimulate muscle action in paralyzed limbs, perhaps allowing paraplegics to walk, is under study. TV dramatizations in which electrical shocks are used to bring a heart attack victim out of ventricular fibrillation (a massively irregular, often fatal, beating of the heart) are more than common. Yet most electrical shock fatalities occur because a current put the heart into fibrillation. A pacemaker uses electrical shocks to stimulate the heart to beat properly. Some fatal shocks do not produce burns, but warts can be safely burned off with electric current (though freezing using liquid nitrogen is now more common). Of course, there are consistent explanations for these disparate effects. The major factors upon which the effects of electrical shock depend are

1. The amount of current I
2. The path taken by the current
3. The duration of the shock
4. The frequency f of the current ($f=0$ for DC)

[link] gives the effects of electrical shocks as a function of current for a typical accidental shock. The effects are for a shock that passes through the trunk of the body, has a duration of 1 s, and is caused by 60-Hz power.



An electric current can cause muscular contractions with varying effects. (a) The victim is "thrown" backward by involuntary muscle contractions that extend the legs and torso. (b) The victim can't let go of the wire that is stimulating all the muscles in the hand. Those that close the fingers are stronger than those that open them.

Current (mA)	**Effect**
1	Threshold of sensation
5	Maximum harmless current
10–20	Onset of sustained muscular contraction; cannot let go for duration of shock; contraction of chest muscles may stop breathing during shock

Current (mA)	Effect
50	Onset of pain
100–300+	Ventricular fibrillation possible; often fatal
300	Onset of burns depending on concentration of current
6000 (6 A)	Onset of sustained ventricular contraction and respiratory paralysis; both cease when shock ends; heartbeat may return to normal; used to defibrillate the heart

Effects of Electrical Shock as a Function of Current[footnote]
For an average male shocked through trunk of body for 1 s by 60-Hz AC. Values for females are 60–80% of those listed.

Our bodies are relatively good conductors due to the water in our bodies. Given that larger currents will flow through sections with lower resistance (to be further discussed in the next chapter), electric currents preferentially flow through paths in the human body that have a minimum resistance in a direct path to earth. The earth is a natural electron sink. Wearing insulating shoes, a requirement in many professions, prohibits a pathway for electrons by providing a large resistance in that path. Whenever working with high-power tools (drills), or in risky situations, ensure that you do not provide a pathway for current flow (especially through the heart).

Very small currents pass harmlessly and unfelt through the body. This happens to you regularly without your knowledge. The threshold of sensation is only 1 mA and, although unpleasant, shocks are apparently harmless for currents less than 5 mA. A great number of safety rules take the 5-mA value for the maximum allowed shock. At 10 to 20 mA and above, the current can stimulate sustained muscular contractions much as regular nerve impulses do. People sometimes say they were knocked across the room by a shock, but what really happened was that certain muscles

contracted, propelling them in a manner not of their own choosing. (See [link](a).) More frightening, and potentially more dangerous, is the “can’t let go” effect illustrated in [link](b). The muscles that close the fingers are stronger than those that open them, so the hand closes involuntarily on the wire shocking it. This can prolong the shock indefinitely. It can also be a danger to a person trying to rescue the victim, because the rescuer’s hand may close about the victim’s wrist. Usually the best way to help the victim is to give the fist a hard knock/blow/jar with an insulator or to throw an insulator at the fist. Modern electric fences, used in animal enclosures, are now pulsed on and off to allow people who touch them to get free, rendering them less lethal than in the past.

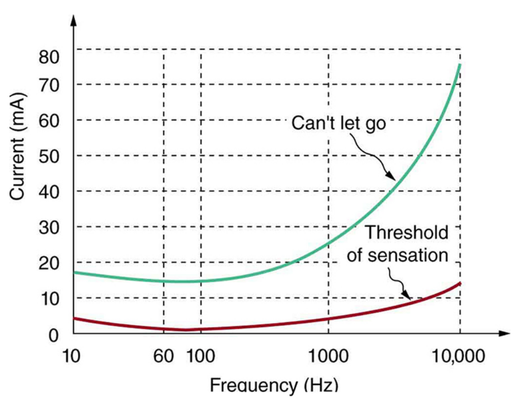
Greater currents may affect the heart. Its electrical patterns can be disrupted, so that it beats irregularly and ineffectively in a condition called “ventricular fibrillation.” This condition often lingers after the shock and is fatal due to a lack of blood circulation. The threshold for ventricular fibrillation is between 100 and 300 mA. At about 300 mA and above, the shock can cause burns, depending on the concentration of current—the more concentrated, the greater the likelihood of burns.

Very large currents cause the heart and diaphragm to contract for the duration of the shock. Both the heart and breathing stop. Interestingly, both often return to normal following the shock. The electrical patterns on the heart are completely erased in a manner that the heart can start afresh with normal beating, as opposed to the permanent disruption caused by smaller currents that can put the heart into ventricular fibrillation. The latter is something like scribbling on a blackboard, whereas the former completely erases it. TV dramatizations of electric shock used to bring a heart attack victim out of ventricular fibrillation also show large paddles. These are used to spread out current passed through the victim to reduce the likelihood of burns.

Current is the major factor determining shock severity (given that other conditions such as path, duration, and frequency are fixed, such as in the table and preceding discussion). A larger voltage is more hazardous, but since $I = \mathrm{V/R}$, the severity of the shock depends on the combination of voltage and resistance. For example, a person with dry skin has a resistance

of about $200\ \mathrm{k\Omega}$. If he comes into contact with 120-V AC, a current $I = (120\ \mathrm{V})/(200\ \mathrm{k\Omega}) = 0.6\ \mathrm{mA}$ passes harmlessly through him. The same person soaking wet may have a resistance of $10.0\ \mathrm{k\Omega}$ and the same 120 V will produce a current of 12 mA—above the “can’t let go” threshold and potentially dangerous.

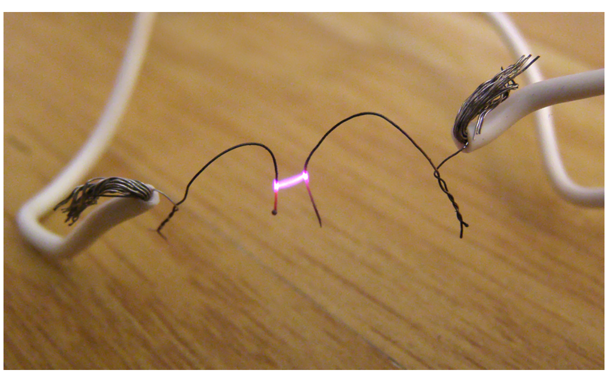
Most of the body’s resistance is in its dry skin. When wet, salts go into ion form, lowering the resistance significantly. The interior of the body has a much lower resistance than dry skin because of all the ionic solutions and fluids it contains. If skin resistance is bypassed, such as by an intravenous infusion, a catheter, or exposed pacemaker leads, a person is rendered **microshock sensitive**. In this condition, currents about 1/1000 those listed in [link] produce similar effects. During open-heart surgery, currents as small as $20\ \mu\mathrm{A}$ can be used to still the heart. Stringent electrical safety requirements in hospitals, particularly in surgery and intensive care, are related to the doubly disadvantaged microshock-sensitive patient. The break in the skin has reduced his resistance, and so the same voltage causes a greater current, and a much smaller current has a greater effect.



Graph of average values for the threshold of sensation and the “can’t let go” current as a function of frequency. The lower the value, the

more sensitive the body is at that frequency.

Factors other than current that affect the severity of a shock are its path, duration, and AC frequency. Path has obvious consequences. For example, the heart is unaffected by an electric shock through the brain, such as may be used to treat manic depression. And it is a general truth that the longer the duration of a shock, the greater its effects. [link] presents a graph that illustrates the effects of frequency on a shock. The curves show the minimum current for two different effects, as a function of frequency. The lower the current needed, the more sensitive the body is at that frequency. Ironically, the body is most sensitive to frequencies near the 50- or 60-Hz frequencies in common use. The body is slightly less sensitive for DC ($f=0$), mildly confirming Edison's claims that AC presents a greater hazard. At higher and higher frequencies, the body becomes progressively less sensitive to any effects that involve nerves. This is related to the maximum rates at which nerves can fire or be stimulated. At very high frequencies, electrical current travels only on the surface of a person. Thus a wart can be burned off with very high frequency current without causing the heart to stop. (Do not try this at home with 60-Hz AC!) Some of the spectacular demonstrations of electricity, in which high-voltage arcs are passed through the air and over people's bodies, employ high frequencies and low currents. (See [link].) Electrical safety devices and techniques are discussed in detail in Electrical Safety: Systems and Devices.



Is this electric arc dangerous?

The answer depends on the AC frequency and the power involved. (credit: Khimich Alex, Wikimedia Commons)

Section Summary

- The two types of electric hazards are thermal (excessive power) and shock (current through a person).
- Shock severity is determined by current, path, duration, and AC frequency.
- [link] lists shock hazards as a function of current.
- [link] graphs the threshold current for two hazards as a function of frequency.

Conceptual Questions

Exercise:

Problem:

Using an ohmmeter, a student measures the resistance between various points on his body. He finds that the resistance between two points on the same finger is about the same as the resistance between two points on opposite hands—both are several hundred thousand ohms. Furthermore, the resistance decreases when more skin is brought into contact with the probes of the ohmmeter. Finally, there is a dramatic drop in resistance (to a few thousand ohms) when the skin is wet. Explain these observations and their implications regarding skin and internal resistance of the human body.

Exercise:

Problem: What are the two major hazards of electricity?

Exercise:

Problem: Why isn't a short circuit a shock hazard?

Exercise:

Problem:

What determines the severity of a shock? Can you say that a certain voltage is hazardous without further information?

Exercise:

Problem:

An electrified needle is used to burn off warts, with the circuit being completed by having the patient sit on a large butt plate. Why is this plate large?

Exercise:

Problem:

Some surgery is performed with high-voltage electricity passing from a metal scalpel through the tissue being cut. Considering the nature of electric fields at the surface of conductors, why would you expect most of the current to flow from the sharp edge of the scalpel? Do you think high- or low-frequency AC is used?

Exercise:

Problem:

Some devices often used in bathrooms, such as hairdryers, often have safety messages saying "Do not use when the bathtub or basin is full of water." Why is this so?

Exercise:

Problem:

We are often advised to not flick electric switches with wet hands, dry your hand first. We are also advised to never throw water on an electric fire. Why is this so?

Exercise:

Problem:

Before working on a power transmission line, linemen will touch the line with the back of the hand as a final check that the voltage is zero. Why the back of the hand?

Exercise:

Problem:

Why is the resistance of wet skin so much smaller than dry, and why do blood and other bodily fluids have low resistances?

Exercise:

Problem:

Could a person on intravenous infusion (an IV) be microshock sensitive?

Exercise:

Problem:

In view of the small currents that cause shock hazards and the larger currents that circuit breakers and fuses interrupt, how do they play a role in preventing shock hazards?

Problem Exercises

Exercise:

Problem:

(a) How much power is dissipated in a short circuit of 240-V AC through a resistance of $0.250\ \Omega$? (b) What current flows?

Solution:

(a) 230 kW

(b) 960 A

Exercise:

Problem:

What voltage is involved in a 1.44-kW short circuit through a $0.100 - \Omega$ resistance?

Exercise:

Problem:

Find the current through a person and identify the likely effect on her if she touches a 120-V AC source: (a) if she is standing on a rubber mat and offers a total resistance of $300\ k\Omega$; (b) if she is standing barefoot on wet grass and has a resistance of only $4000\ k\Omega$.

Solution:

(a) 0.400 mA, no effect

(b) 26.7 mA, muscular contraction for duration of the shock (can't let go)

Exercise:

Problem:

While taking a bath, a person touches the metal case of a radio. The path through the person to the drainpipe and ground has a resistance of $4000\ \Omega$. What is the smallest voltage on the case of the radio that could cause ventricular fibrillation?

Exercise:

Problem:

Foolishly trying to fish a burning piece of bread from a toaster with a metal butter knife, a man comes into contact with 120-V AC. He does not even feel it since, luckily, he is wearing rubber-soled shoes. What is the minimum resistance of the path the current follows through the person?

Solution:

$1.20 \times 10^5\ \Omega$

Exercise:

Problem:

(a) During surgery, a current as small as $20.0\ \mu A$ applied directly to the heart may cause ventricular fibrillation. If the resistance of the exposed heart is $300\ \Omega$, what is the smallest voltage that poses this danger? (b) Does your answer imply that special electrical safety precautions are needed?

Exercise:

Problem:

(a) What is the resistance of a 220-V AC short circuit that generates a peak power of 96.8 kW? (b) What would the average power be if the voltage was 120 V AC?

Solution:

(a) $1.00\ \Omega$

(b) 14.4 kW

Exercise:

Problem:

A heart defibrillator passes 10.0 A through a patient’s torso for 5.00 ms in an attempt to restore normal beating. (a) How much charge passed? (b) What voltage was applied if 500 J of energy was dissipated? (c) What was the path’s resistance? (d) Find the temperature increase caused in the 8.00 kg of affected tissue.

Exercise:

Problem: Integrated Concepts

A short circuit in a 120-V appliance cord has a $0.500\text{-}\Omega$ resistance. Calculate the temperature rise of the 2.00 g of surrounding materials, assuming their specific heat capacity is $0.200\ \text{cal/g}\cdot^{\circ}\text{C}$ and that it takes 0.0500 s for a circuit breaker to interrupt the current. Is this likely to be damaging?

Solution:

Temperature increases 860° C. It is very likely to be damaging.

Exercise:

Problem: Construct Your Own Problem

Consider a person working in an environment where electric currents might pass through her body. Construct a problem in which you calculate the resistance of insulation needed to protect the person from harm. Among the things to be considered are the voltage to which the person might be exposed, likely body resistance (dry, wet, …), and acceptable currents (safe but sensed, safe and unfelt, …).

Glossary

thermal hazard
: a hazard in which electric current causes undesired thermal effects

shock hazard

when electric current passes through a person

short circuit

also known as a “short,” a low-resistance path between terminals of a voltage source

microshock sensitive

a condition in which a person’s skin resistance is bypassed, possibly by a medical procedure, rendering the person vulnerable to electrical shock at currents about 1/1000 the normally required level

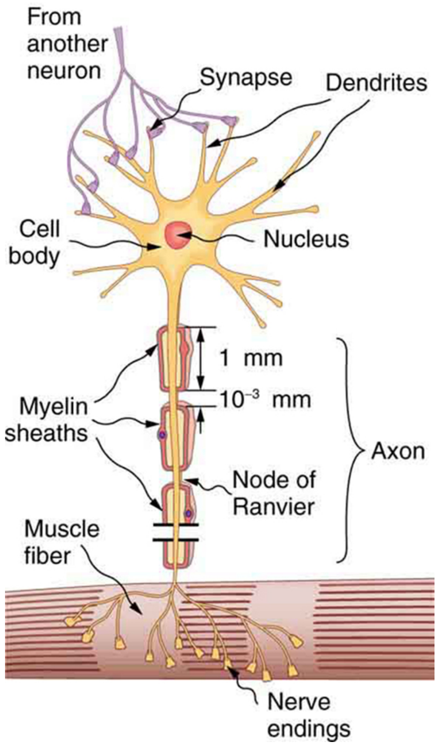
Nerve Conduction–Electrocardiograms

- Explain the process by which electric signals are transmitted along a neuron.
- Explain the effects myelin sheaths have on signal propagation.
- Explain what the features of an ECG signal indicate.

Nerve Conduction

Electric currents in the vastly complex system of billions of nerves in our body allow us to sense the world, control parts of our body, and think. These are representative of the three major functions of nerves. First, nerves carry messages from our sensory organs and others to the central nervous system, consisting of the brain and spinal cord. Second, nerves carry messages from the central nervous system to muscles and other organs. Third, nerves transmit and process signals within the central nervous system. The sheer number of nerve cells and the incredibly greater number of connections between them makes this system the subtle wonder that it is. **Nerve conduction** is a general term for electrical signals carried by nerve cells. It is one aspect of **bioelectricity**, or electrical effects in and created by biological systems.

Nerve cells, properly called *neurons*, look different from other cells—they have tendrils, some of them many centimeters long, connecting them with other cells. (See [link].) Signals arrive at the cell body across *synapses* or through *dendrites*, stimulating the neuron to generate its own signal, sent along its long *axon* to other nerve or muscle cells. Signals may arrive from many other locations and be transmitted to yet others, conditioning the synapses by use, giving the system its complexity and its ability to learn.

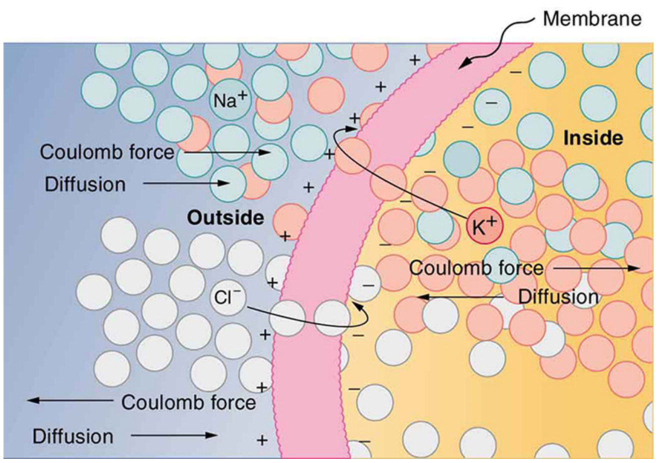


A neuron with its dendrites and long axon. Signals in the form of electric currents reach the cell body through dendrites and across synapses, stimulating the neuron to generate its own signal sent down the axon. The number of interconnections can be far greater than shown here.

The method by which these electric currents are generated and transmitted is more complex than the simple movement of free charges in a conductor,

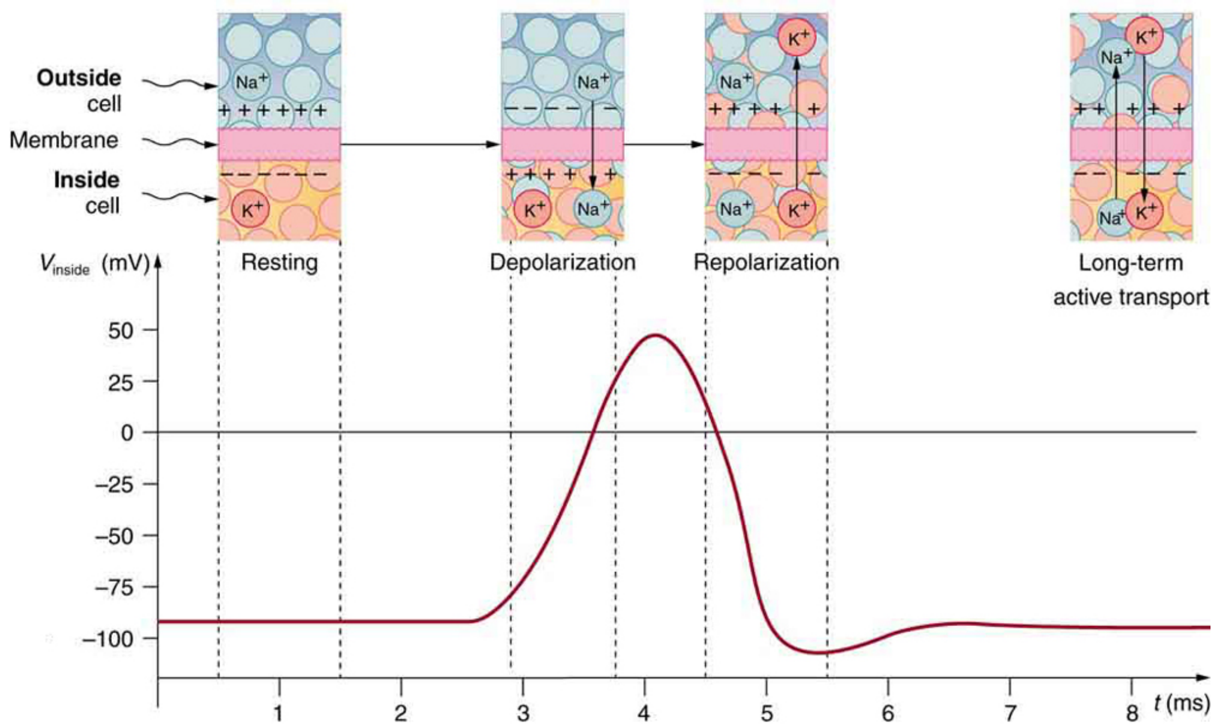
but it can be understood with principles already discussed in this text. The most important of these are the Coulomb force and diffusion.

[link] illustrates how a voltage (potential difference) is created across the cell membrane of a neuron in its resting state. This thin membrane separates electrically neutral fluids having differing concentrations of ions, the most important varieties being Na^+, K^+, and Cl^- (these are sodium, potassium, and chlorine ions with single plus or minus charges as indicated). As discussed in Molecular Transport Phenomena: Diffusion, Osmosis, and Related Processes, free ions will diffuse from a region of high concentration to one of low concentration. But the cell membrane is **semipermeable**, meaning that some ions may cross it while others cannot. In its resting state, the cell membrane is permeable to K^+ and Cl^-, and impermeable to Na^+. Diffusion of K^+ and Cl^- thus creates the layers of positive and negative charge on the outside and inside of the membrane. The Coulomb force prevents the ions from diffusing across in their entirety. Once the charge layer has built up, the repulsion of like charges prevents more from moving across, and the attraction of unlike charges prevents more from leaving either side. The result is two layers of charge right on the membrane, with diffusion being balanced by the Coulomb force. A tiny fraction of the charges move across and the fluids remain neutral (other ions are present), while a separation of charge and a voltage have been created across the membrane.



The semipermeable membrane of a

cell has different concentrations of ions inside and out. Diffusion moves the K^+ and Cl^- ions in the direction shown, until the Coulomb force halts further transfer. This results in a layer of positive charge on the outside, a layer of negative charge on the inside, and thus a voltage across the cell membrane. The membrane is normally impermeable to Na^+.



An action potential is the pulse of voltage inside a nerve cell graphed here. It is caused by movements of ions across the cell membrane as shown. Depolarization occurs when a stimulus makes the membrane permeable to Na^+ ions. Repolarization follows as the membrane

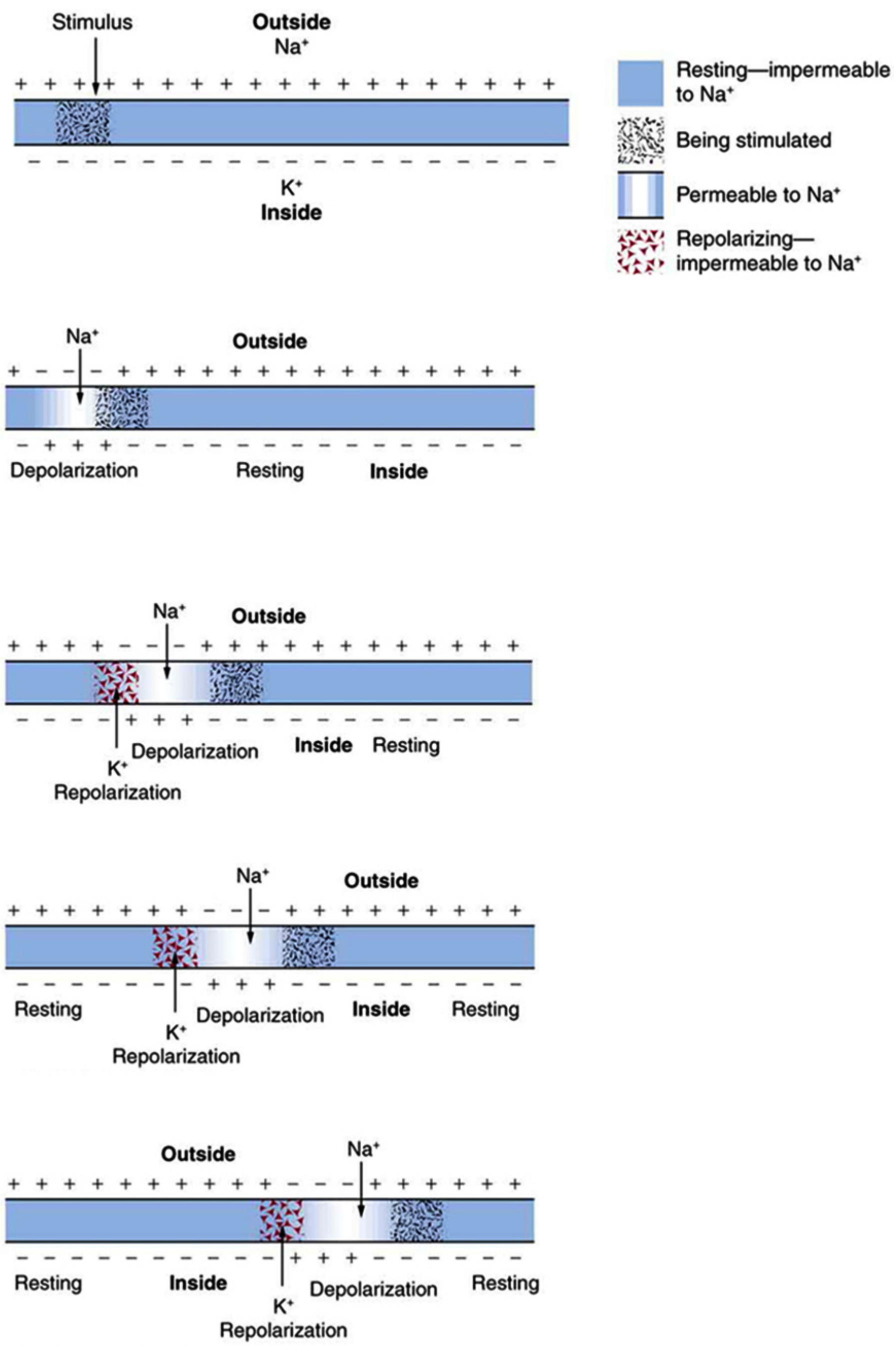
again becomes impermeable to Na^+, and K^+ moves from high to low concentration. In the long term, active transport slowly maintains the concentration differences, but the cell may fire hundreds of times in rapid succession without seriously depleting them.

The separation of charge creates a potential difference of 70 to 90 mV across the cell membrane. While this is a small voltage, the resulting electric field ($E = V/d$) across the only 8-nm-thick membrane is immense (on the order of 11 MV/m!) and has fundamental effects on its structure and permeability. Now, if the exterior of a neuron is taken to be at 0 V, then the interior has a *resting potential* of about –90 mV. Such voltages are created across the membranes of almost all types of animal cells but are largest in nerve and muscle cells. In fact, fully 25% of the energy used by cells goes toward creating and maintaining these potentials.

Electric currents along the cell membrane are created by any stimulus that changes the membrane's permeability. The membrane thus temporarily becomes permeable to Na^+, which then rushes in, driven both by diffusion and the Coulomb force. This inrush of Na^+ first neutralizes the inside membrane, or *depolarizes* it, and then makes it slightly positive. The depolarization causes the membrane to again become impermeable to Na^+, and the movement of K^+ quickly returns the cell to its resting potential, or *repolarizes* it. This sequence of events results in a voltage pulse, called the *action potential*. (See [link].) Only small fractions of the ions move, so that the cell can fire many hundreds of times without depleting the excess concentrations of Na^+ and K^+. Eventually, the cell must replenish these ions to maintain the concentration differences that create bioelectricity. This sodium-potassium pump is an example of *active transport*, wherein cell energy is used to move ions across membranes against diffusion gradients and the Coulomb force.

The action potential is a voltage pulse at one location on a cell membrane. How does it get transmitted along the cell membrane, and in particular down an axon, as a nerve impulse? The answer is that the changing voltage and electric fields affect the permeability of the adjacent cell membrane, so

that the same process takes place there. The adjacent membrane depolarizes, affecting the membrane further down, and so on, as illustrated in [link]. Thus the action potential stimulated at one location triggers a *nerve impulse* that moves slowly (about 1 m/s) along the cell membrane.



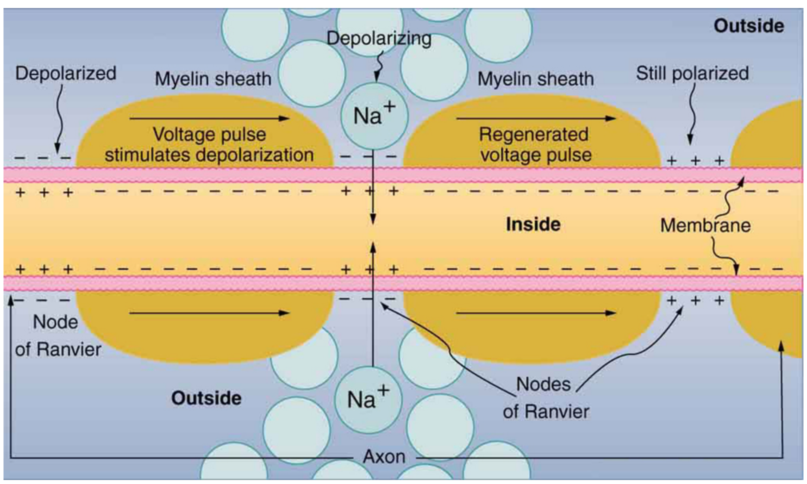
A nerve impulse is the propagation of an action potential along a cell membrane. A stimulus causes an action potential at one location, which changes the permeability of the adjacent membrane, causing an action potential there. This in turn affects the membrane further down, so that the action potential moves slowly (in electrical terms) along the cell membrane. Although the impulse is due to Na^{+} and K^{+} going across the membrane, it is equivalent to a wave of charge moving along the outside and inside of the membrane.

Some axons, like that in [link], are sheathed with *myelin*, consisting of fat-containing cells. [link] shows an enlarged view of an axon having myelin sheaths characteristically separated by unmyelinated gaps (called nodes of Ranvier). This arrangement gives the axon a number of interesting properties. Since myelin is an insulator, it prevents signals from jumping between adjacent nerves (cross talk). Additionally, the myelinated regions transmit electrical signals at a very high speed, as an ordinary conductor or resistor would. There is no action potential in the myelinated regions, so that no cell energy is used in them. There is an IR signal loss in the myelin, but the signal is regenerated in the gaps, where the voltage pulse triggers the action potential at full voltage. So a myelinated axon transmits a nerve impulse faster, with less energy consumption, and is better protected from cross talk than an unmyelinated one. Not all axons are myelinated, so that cross talk and slow signal transmission are a characteristic of the normal operation of these axons, another variable in the nervous system.

The degeneration or destruction of the myelin sheaths that surround the nerve fibers impairs signal transmission and can lead to numerous neurological effects. One of the most prominent of these diseases comes from the body’s own immune system attacking the myelin in the central nervous system—multiple sclerosis. MS symptoms include fatigue, vision problems, weakness of arms and legs, loss of balance, and tingling or

numbness in one’s extremities (neuropathy). It is more apt to strike younger adults, especially females. Causes might come from infection, environmental or geographic affects, or genetics. At the moment there is no known cure for MS.

Most animal cells can fire or create their own action potential. Muscle cells contract when they fire and are often induced to do so by a nerve impulse. In fact, nerve and muscle cells are physiologically similar, and there are even hybrid cells, such as in the heart, that have characteristics of both nerves and muscles. Some animals, like the infamous electric eel (see [link]), use muscles ganged so that their voltages add in order to create a shock great enough to stun prey.

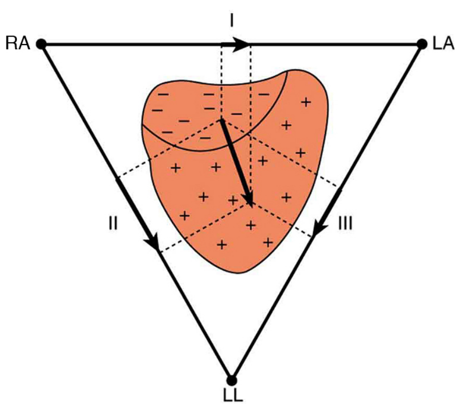


Propagation of a nerve impulse down a myelinated axon, from left to right. The signal travels very fast and without energy input in the myelinated regions, but it loses voltage. It is regenerated in the gaps. The signal moves faster than in unmyelinated axons and is insulated from signals in other nerves, limiting cross talk.

An electric eel flexes its muscles to create a voltage that stuns prey. (credit: chrisbb, Flickr)

Electrocardiograms

Just as nerve impulses are transmitted by depolarization and repolarization of adjacent membrane, the depolarization that causes muscle contraction can also stimulate adjacent muscle cells to depolarize (fire) and contract. Thus, a depolarization wave can be sent across the heart, coordinating its rhythmic contractions and enabling it to perform its vital function of propelling blood through the circulatory system. [link] is a simplified graphic of a depolarization wave spreading across the heart from the *sinoarterial (SA) node***,** the heart's natural pacemaker.

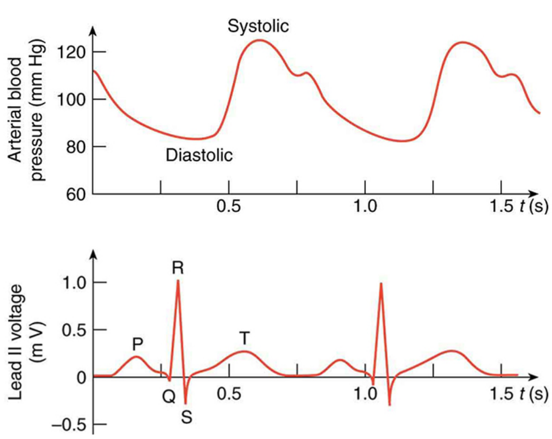


The outer surface of the heart changes from positive to negative during depolarization. This wave of depolarization is spreading from the top of the heart and is represented by a vector pointing in the direction of the wave. This vector is a voltage (potential difference) vector. Three electrodes, labeled RA, LA, and LL, are placed on the patient. Each pair (called leads I, II, and III) measures a component of the depolarization vector and is graphed in an ECG.

An **electrocardiogram (ECG)** is a record of the voltages created by the wave of depolarization and subsequent repolarization in the heart. Voltages between pairs of electrodes placed on the chest are vector components of the voltage wave on the heart. Standard ECGs have 12 or more electrodes, but only three are shown in [link] for clarity. Decades ago, three-electrode ECGs were performed by placing electrodes on the left and right arms and the left leg. The voltage between the right arm and the left leg is called the *lead II potential* and is the most often graphed. We shall examine the lead II potential as an indicator of heart-muscle function and see that it is coordinated with arterial blood pressure as well.

Heart function and its four-chamber action are explored in [Viscosity and Laminar Flow; Poiseuille's Law](). Basically, the right and left atria receive blood from the body and lungs, respectively, and pump the blood into the ventricles. The right and left ventricles, in turn, pump blood through the lungs and the rest of the body, respectively. Depolarization of the heart muscle causes it to contract. After contraction it is repolarized to ready it for the next beat. The ECG measures components of depolarization and repolarization of the heart muscle and can yield significant information on the functioning and malfunctioning of the heart.

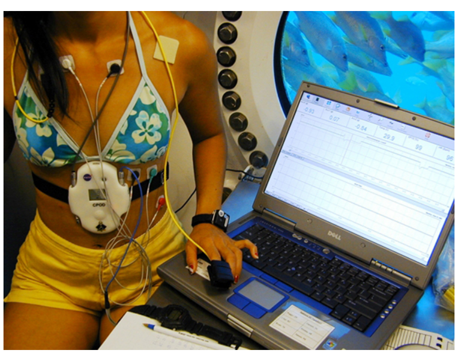
[link] shows an ECG of the lead II potential and a graph of the corresponding arterial blood pressure. The major features are labeled P, Q, R, S, and T. The *P wave* is generated by the depolarization and contraction of the atria as they pump blood into the ventricles. The *QRS complex* is created by the depolarization of the ventricles as they pump blood to the lungs and body. Since the shape of the heart and the path of the depolarization wave are not simple, the QRS complex has this typical shape and time span. The lead II QRS signal also masks the repolarization of the atria, which occur at the same time. Finally, the *T wave* is generated by the repolarization of the ventricles and is followed by the next P wave in the next heartbeat. Arterial blood pressure varies with each part of the heartbeat, with systolic (maximum) pressure occurring closely after the QRS complex, which signals contraction of the ventricles.



A lead II ECG with

corresponding arterial blood pressure. The QRS complex is created by the depolarization and contraction of the ventricles and is followed shortly by the maximum or systolic blood pressure. See text for further description.

Taken together, the 12 leads of a state-of-the-art ECG can yield a wealth of information about the heart. For example, regions of damaged heart tissue, called infarcts, reflect electrical waves and are apparent in one or more lead potentials. Subtle changes due to slight or gradual damage to the heart are most readily detected by comparing a recent ECG to an older one. This is particularly the case since individual heart shape, size, and orientation can cause variations in ECGs from one individual to another. ECG technology has advanced to the point where a portable ECG monitor with a liquid crystal instant display and a printer can be carried to patients' homes or used in emergency vehicles. See [link].



This NASA scientist and NEEMO 5 aquanaut's heart rate and other vital signs

are being recorded by a portable device while living in an underwater habitat. (credit: NASA, Life Sciences Data Archive at Johnson Space Center, Houston, Texas)

Note:
PhET Explorations: Neuron
Stimulate a neuron and monitor what happens. Pause, rewind, and move forward in time in order to observe the ions as they move across the neuron membrane.
https://phet.colorado.edu/sims/html/neuron/latest/neuron_en.html

Section Summary

- Electric potentials in neurons and other cells are created by ionic concentration differences across semipermeable membranes.
- Stimuli change the permeability and create action potentials that propagate along neurons.
- Myelin sheaths speed this process and reduce the needed energy input.
- This process in the heart can be measured with an electrocardiogram (ECG).

Conceptual Questions

Exercise:

Problem:

Note that in [link], both the concentration gradient and the Coulomb force tend to move Na^+ ions into the cell. What prevents this?

Exercise:

Problem:

Define depolarization, repolarization, and the action potential.

Exercise:

Problem:

Explain the properties of myelinated nerves in terms of the insulating properties of myelin.

Problems & Exercises

Exercise:

Problem: Integrated Concepts

Use the ECG in [link] to determine the heart rate in beats per minute assuming a constant time between beats.

Solution:

80 beats/minute

Exercise:

Problem: Integrated Concepts

(a) Referring to [link], find the time systolic pressure lags behind the middle of the QRS complex. (b) Discuss the reasons for the time lag.

Glossary

nerve conduction
the transport of electrical signals by nerve cells

bioelectricity
electrical effects in and created by biological systems

semipermeable
property of a membrane that allows only certain types of ions to cross it

electrocardiogram (ECG)
usually abbreviated ECG, a record of voltages created by depolarization and repolarization, especially in the heart

Introduction to Circuits and DC Instruments
class="introduction"

Electric circuits in a computer allow large amounts of data to be quickly and accurately analyzed.. (credit: Airman 1st Class Mike Meares, United States Air Force)

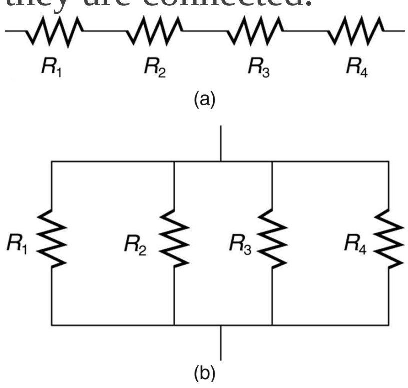
Electric circuits are commonplace. Some are simple, such as those in flashlights. Others, such as those used in supercomputers, are extremely complex.

This collection of modules takes the topic of electric circuits a step beyond simple circuits. When the circuit is purely resistive, everything in this module applies to both DC and AC. Matters become more complex when capacitance is involved. We do consider what happens when capacitors are connected to DC voltage sources, but the interaction of capacitors and other nonresistive devices with AC is left for a later chapter. Finally, a number of important DC instruments, such as meters that measure voltage and current, are covered in this chapter.

Resistors in Series and Parallel

- Draw a circuit with resistors in parallel and in series.
- Calculate the voltage drop of a current across a resistor using Ohm's law.
- Contrast the way total resistance is calculated for resistors in series and in parallel.
- Explain why total resistance of a parallel circuit is less than the smallest resistance of any of the resistors in that circuit.
- Calculate total resistance of a circuit that contains a mixture of resistors connected in series and in parallel.

Most circuits have more than one component, called a **resistor** that limits the flow of charge in the circuit. A measure of this limit on charge flow is called **resistance**. The simplest combinations of resistors are the series and parallel connections illustrated in [link]. The total resistance of a combination of resistors depends on both their individual values and how they are connected.

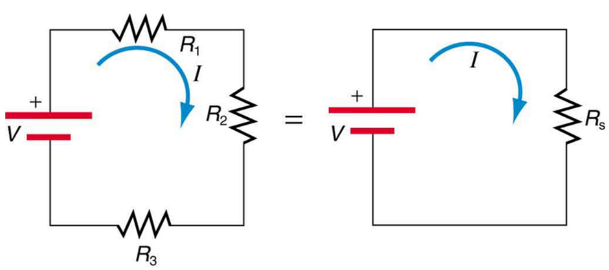


(a) A series connection of resistors. (b) A parallel connection of resistors.

Resistors in Series

When are resistors in **series**? Resistors are in series whenever the flow of charge, called the **current**, must flow through devices sequentially. For example, if current flows through a person holding a screwdriver and into the Earth, then R_1 in [link](a) could be the resistance of the screwdriver's shaft, R_2 the resistance of its handle, R_3 the person's body resistance, and R_4 the resistance of her shoes.

[link] shows resistors in series connected to a **voltage** source. It seems reasonable that the total resistance is the sum of the individual resistances, considering that the current has to pass through each resistor in sequence. (This fact would be an advantage to a person wishing to avoid an electrical shock, who could reduce the current by wearing high-resistance rubber-soled shoes. It could be a disadvantage if one of the resistances were a faulty high-resistance cord to an appliance that would reduce the operating current.)



Three resistors connected in series to a battery (left) and the equivalent single or series resistance (right).

To verify that resistances in series do indeed add, let us consider the loss of electrical power, called a **voltage drop**, in each resistor in [link].

According to **Ohm's law**, the voltage drop, V, across a resistor when a current flows through it is calculated using the equation $V = \mathrm{IR}$, where I equals the current in amps (A) and R is the resistance in ohms (Ω). Another

way to think of this is that V is the voltage necessary to make a current I flow through a resistance R.

So the voltage drop across R_1 is $V_1 = IR_1$, that across R_2 is $V_2 = IR_2$, and that across R_3 is $V_3 = IR_3$. The sum of these voltages equals the voltage output of the source; that is,

Equation:

$$V = V_1 + V_2 + V_3.$$

This equation is based on the conservation of energy and conservation of charge. Electrical potential energy can be described by the equation $\mathrm{PE} = \mathrm{qV}$, where q is the electric charge and V is the voltage. Thus the energy supplied by the source is qV, while that dissipated by the resistors is

Equation:

$$qV_1 + qV_2 + qV_3.$$

Note:

Connections: Conservation Laws

The derivations of the expressions for series and parallel resistance are based on the laws of conservation of energy and conservation of charge, which state that total charge and total energy are constant in any process. These two laws are directly involved in all electrical phenomena and will be invoked repeatedly to explain both specific effects and the general behavior of electricity.

These energies must be equal, because there is no other source and no other destination for energy in the circuit. Thus, $qV = qV_1 + qV_2 + qV_3$. The charge q cancels, yielding $V = V_1 + V_2 + V_3$, as stated. (Note that the same amount of charge passes through the battery and each resistor in a given amount of time, since there is no capacitance to store charge, there is no place for charge to leak, and charge is conserved.)

Now substituting the values for the individual voltages gives
Equation:

$$V = IR_1 + IR_2 + IR_3 = I(R_1 + R_2 + R_3).$$

Note that for the equivalent single series resistance R_s, we have
Equation:

$$V = IR_s.$$

This implies that the total or equivalent series resistance R_s of three resistors is $R_s = R_1 + R_2 + R_3$.

This logic is valid in general for any number of resistors in series; thus, the total resistance R_s of a series connection is
Equation:

$$R_s = R_1 + R_2 + R_3 + ...,$$

as proposed. Since all of the current must pass through each resistor, it experiences the resistance of each, and resistances in series simply add up.

Example:
Calculating Resistance, Current, Voltage Drop, and Power Dissipation: Analysis of a Series Circuit
Suppose the voltage output of the battery in [link] is $12.0\ \text{V}$, and the resistances are $R_1 = 1.00\ \Omega$, $R_2 = 6.00\ \Omega$, and $R_3 = 13.0\ \Omega$. (a) What is the total resistance? (b) Find the current. (c) Calculate the voltage drop in each resistor, and show these add to equal the voltage output of the source. (d) Calculate the power dissipated by each resistor. (e) Find the power output of the source, and show that it equals the total power dissipated by the resistors.
Strategy and Solution for (a)

The total resistance is simply the sum of the individual resistances, as given by this equation:

Equation:

$$\begin{aligned} R_s &= R_1 + R_2 + R_3 \\ &= 1.00\ \Omega + 6.00\ \Omega + 13.0\ \Omega \\ &= 20.0\ \Omega. \end{aligned}$$

Strategy and Solution for (b)

The current is found using Ohm's law, $V = \text{IR}$. Entering the value of the applied voltage and the total resistance yields the current for the circuit:

Equation:

$$I = \frac{V}{R_s} = \frac{12.0\ \text{V}}{20.0\ \Omega} = 0.600\ \text{A}.$$

Strategy and Solution for (c)

The voltage—or IR drop—in a resistor is given by Ohm's law. Entering the current and the value of the first resistance yields

Equation:

$$V_1 = IR_1 = (0.600\ \text{A})(1.0\ \Omega) = 0.600\ \text{V}.$$

Similarly,

Equation:

$$V_2 = IR_2 = (0.600\ \text{A})(6.0\ \Omega) = 3.60\ \text{V}$$

and

Equation:

$$V_3 = IR_3 = (0.600\ \text{A})(13.0\ \Omega) = 7.80\ \text{V}.$$

Discussion for (c)

The three IR drops add to $12.0\ \text{V}$, as predicted:

Equation:

$$V_1 + V_2 + V_3 = (0.600 + 3.60 + 7.80)\ \text{V} = 12.0\ \text{V}.$$

Strategy and Solution for (d)

The easiest way to calculate power in watts (W) dissipated by a resistor in a DC circuit is to use **Joule's law**, $P = \text{IV}$, where P is electric power. In this case, each resistor has the same full current flowing through it. By substituting Ohm's law $V = \text{IR}$ into Joule's law, we get the power dissipated by the first resistor as

Equation:

$$P_1 = I^2 R_1 = (0.600 \text{ A})^2 (1.00 \ \Omega) = 0.360 \text{ W}.$$

Similarly,

Equation:

$$P_2 = I^2 R_2 = (0.600 \text{ A})^2 (6.00 \ \Omega) = 2.16 \text{ W}$$

and

Equation:

$$P_3 = I^2 R_3 = (0.600 \text{ A})^2 (13.0 \ \Omega) = 4.68 \text{ W}.$$

Discussion for (d)

Power can also be calculated using either $P = \text{IV}$ or $P = \frac{V^2}{R}$, where V is the voltage drop across the resistor (not the full voltage of the source). The same values will be obtained.

Strategy and Solution for (e)

The easiest way to calculate power output of the source is to use $P = \text{IV}$, where V is the source voltage. This gives

Equation:

$$P = (0.600 \text{ A})(12.0 \text{ V}) = 7.20 \text{ W}.$$

Discussion for (e)

Note, coincidentally, that the total power dissipated by the resistors is also 7.20 W, the same as the power put out by the source. That is,

Equation:

$$P_1 + P_2 + P_3 = (0.360 + 2.16 + 4.68) \text{ W} = 7.20 \text{ W}.$$

Power is energy per unit time (watts), and so conservation of energy requires the power output of the source to be equal to the total power dissipated by the resistors.

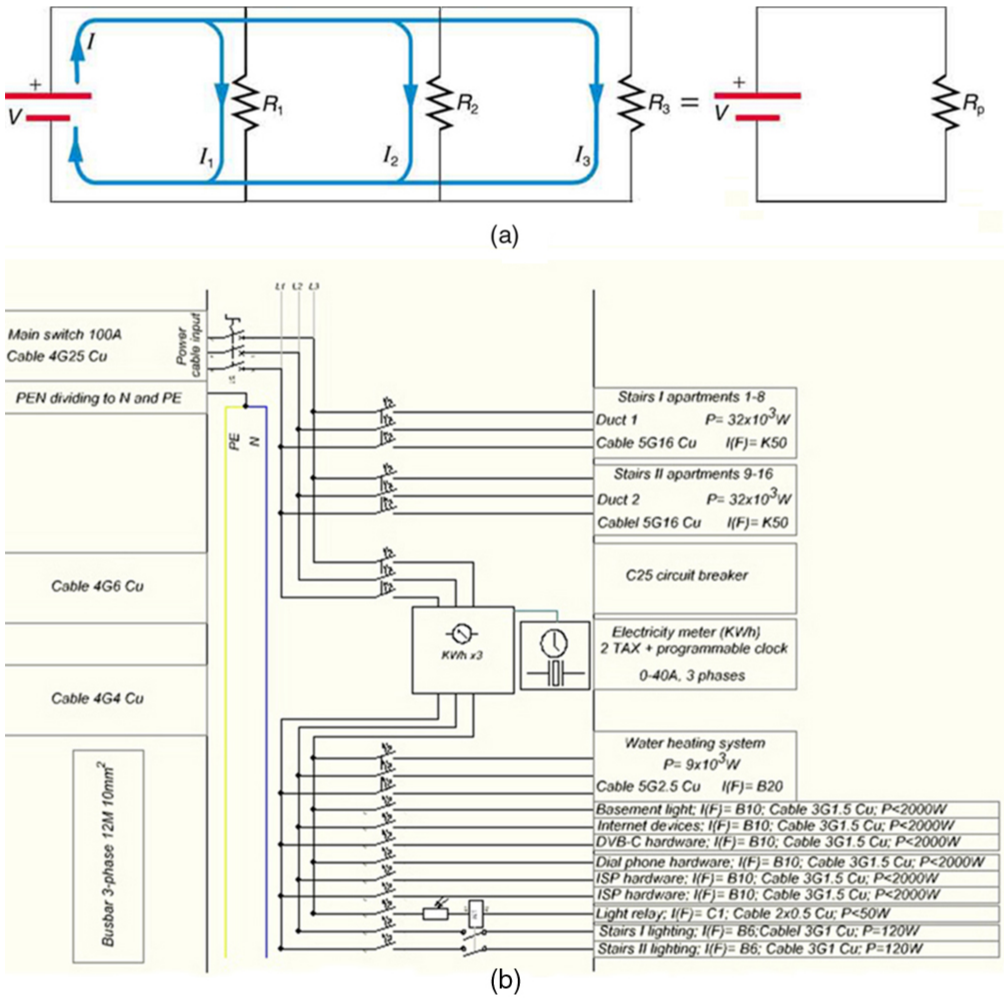
Note:
Major Features of Resistors in Series

1. Series resistances add: $R_s = R_1 + R_2 + R_3 + ...$
2. The same current flows through each resistor in series.
3. Individual resistors in series do not get the total source voltage, but divide it.

Resistors in Parallel

[link] shows resistors in **parallel**, wired to a voltage source. Resistors are in parallel when each resistor is connected directly to the voltage source by connecting wires having negligible resistance. Each resistor thus has the full voltage of the source applied to it.

Each resistor draws the same current it would if it alone were connected to the voltage source (provided the voltage source is not overloaded). For example, an automobile's headlights, radio, and so on, are wired in parallel, so that they utilize the full voltage of the source and can operate completely independently. The same is true in your house, or any building. (See [link] (b).)



(a) Three resistors connected in parallel to a battery and the equivalent single or parallel resistance. (b) Electrical power setup in a house. (credit: Dmitry G, Wikimedia Commons)

To find an expression for the equivalent parallel resistance R_{p}, let us consider the currents that flow and how they are related to resistance. Since each resistor in the circuit has the full voltage, the currents flowing through the individual resistors are $I_1 = \frac{V}{R_1}$, $I_2 = \frac{V}{R_2}$, and $I_3 = \frac{V}{R_3}$. Conservation of charge implies that the total current I produced by the source is the sum of these currents:

Equation:

$$I = I_1 + I_2 + I_3.$$

Substituting the expressions for the individual currents gives

Equation:

$$I = \frac{V}{R_1} + \frac{V}{R_2} + \frac{V}{R_3} = V\left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right).$$

Note that Ohm's law for the equivalent single resistance gives

Equation:

$$I = \frac{V}{R_\text{p}} = V\left(\frac{1}{R_\text{p}}\right).$$

The terms inside the parentheses in the last two equations must be equal. Generalizing to any number of resistors, the total resistance R_p of a parallel connection is related to the individual resistances by

Equation:

$$\frac{1}{R_\text{p}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_{.3}} + \ldots.$$

This relationship results in a total resistance R_p that is less than the smallest of the individual resistances. (This is seen in the next example.) When resistors are connected in parallel, more current flows from the source than would flow for any of them individually, and so the total resistance is lower.

Example:

Calculating Resistance, Current, Power Dissipation, and Power Output: Analysis of a Parallel Circuit

Let the voltage output of the battery and resistances in the parallel connection in [link] be the same as the previously considered series

connection: $V = 12.0\text{ V}$, $R_1 = 1.00\ \Omega$, $R_2 = 6.00\ \Omega$, and $R_3 = 13.0\ \Omega$. (a) What is the total resistance? (b) Find the total current. (c) Calculate the currents in each resistor, and show these add to equal the total current output of the source. (d) Calculate the power dissipated by each resistor. (e) Find the power output of the source, and show that it equals the total power dissipated by the resistors.

Strategy and Solution for (a)

The total resistance for a parallel combination of resistors is found using the equation below. Entering known values gives

Equation:

$$\frac{1}{R_\mathrm{p}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} = \frac{1}{1.00\ \Omega} + \frac{1}{6.00\ \Omega} + \frac{1}{13.0\ \Omega}.$$

Thus,

Equation:

$$\frac{1}{R_\mathrm{p}} = \frac{1.00}{\Omega} + \frac{0.1667}{\Omega} + \frac{0.07692}{\Omega} = \frac{1.2436}{\Omega}.$$

(Note that in these calculations, each intermediate answer is shown with an extra digit.)

We must invert this to find the total resistance R_p. This yields

Equation:

$$R_\mathrm{p} = \frac{1}{1.2436}\Omega = 0.8041\ \Omega.$$

The total resistance with the correct number of significant digits is $R_\mathrm{p} = 0.804\ \Omega$.

Discussion for (a)

R_p is, as predicted, less than the smallest individual resistance.

Strategy and Solution for (b)

The total current can be found from Ohm's law, substituting R_p for the total resistance. This gives

Equation:

$$I = \frac{V}{R_{\mathrm{p}}} = \frac{12.0\ \mathrm{V}}{0.8041\ \Omega} = 14.92\ \mathrm{A}.$$

Discussion for (b)
Current I for each device is much larger than for the same devices connected in series (see the previous example). A circuit with parallel connections has a smaller total resistance than the resistors connected in series.
Strategy and Solution for (c)
The individual currents are easily calculated from Ohm's law, since each resistor gets the full voltage. Thus,
Equation:

$$I_1 = \frac{V}{R_1} = \frac{12.0\ \mathrm{V}}{1.00\,\Omega} = 12.0\ \mathrm{A}.$$

Similarly,
Equation:

$$I_2 = \frac{V}{R_2} = \frac{12.0\ \mathrm{V}}{6.00\,\Omega} = 2.00\ \mathrm{A}$$

and
Equation:

$$I_3 = \frac{V}{R_3} = \frac{12.0\ \mathrm{V}}{13.0\,\Omega} = 0.92\ \mathrm{A}.$$

Discussion for (c)
The total current is the sum of the individual currents:
Equation:

$$I_1 + I_2 + I_3 = 14.92\ \mathrm{A}.$$

This is consistent with conservation of charge.
Strategy and Solution for (d)
The power dissipated by each resistor can be found using any of the equations relating power to current, voltage, and resistance, since all three

are known. Let us use $P = \frac{V^2}{R}$, since each resistor gets full voltage. Thus,

Equation:

$$P_1 = \frac{V^2}{R_1} = \frac{(12.0\text{ V})^2}{1.00\ \Omega} = 144\text{ W}.$$

Similarly,

Equation:

$$P_2 = \frac{V^2}{R_2} = \frac{(12.0\text{ V})^2}{6.00\ \Omega} = 24.0\text{ W}$$

and

Equation:

$$P_3 = \frac{V^2}{R_3} = \frac{(12.0\text{ V})^2}{13.0\ \Omega} = 11.1\text{ W}.$$

Discussion for (d)

The power dissipated by each resistor is considerably higher in parallel than when connected in series to the same voltage source.

Strategy and Solution for (e)

The total power can also be calculated in several ways. Choosing $P = \text{IV}$, and entering the total current, yields

Equation:

$$P = \text{IV} = (14.92\text{ A})(12.0\text{ V}) = 179\text{ W}.$$

Discussion for (e)

Total power dissipated by the resistors is also 179 W:

Equation:

$$P_1 + P_2 + P_3 = 144\text{ W} + 24.0\text{ W} + 11.1\text{ W} = 179\text{ W}.$$

This is consistent with the law of conservation of energy.

Overall Discussion

Note that both the currents and powers in parallel connections are greater than for the same devices in series.

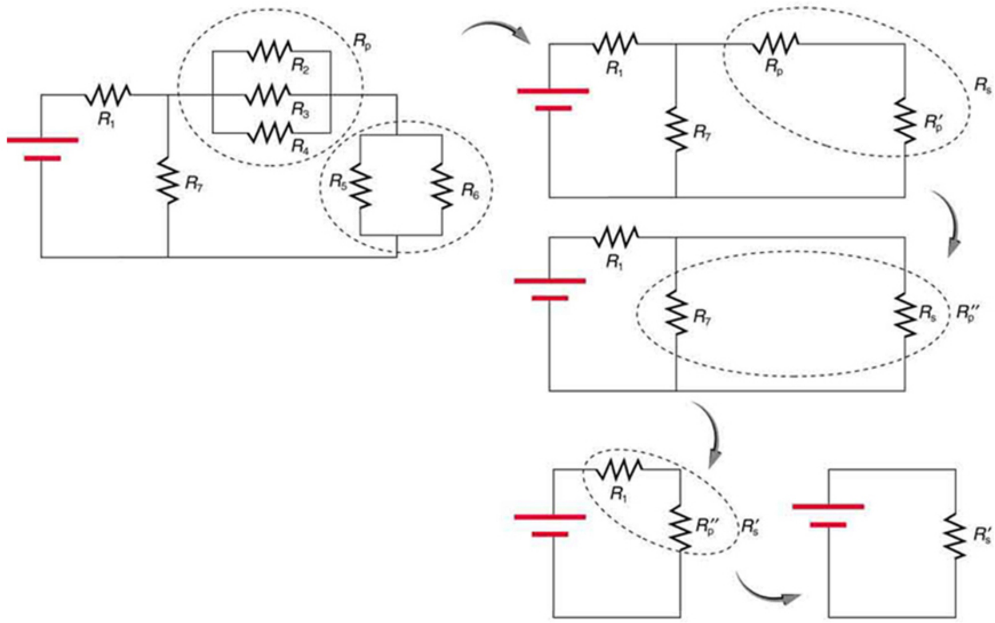
Note:
Major Features of Resistors in Parallel

1. Parallel resistance is found from $\frac{1}{R_{\text{p}}}=\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3}+...$, and it is smaller than any individual resistance in the combination.
2. Each resistor in parallel has the same full voltage of the source applied to it. (Power distribution systems most often use parallel connections to supply the myriad devices served with the same voltage and to allow them to operate independently.)
3. Parallel resistors do not each get the total current; they divide it.

Combinations of Series and Parallel

More complex connections of resistors are sometimes just combinations of series and parallel. These are commonly encountered, especially when wire resistance is considered. In that case, wire resistance is in series with other resistances that are in parallel.

Combinations of series and parallel can be reduced to a single equivalent resistance using the technique illustrated in [link]. Various parts are identified as either series or parallel, reduced to their equivalents, and further reduced until a single resistance is left. The process is more time consuming than difficult.

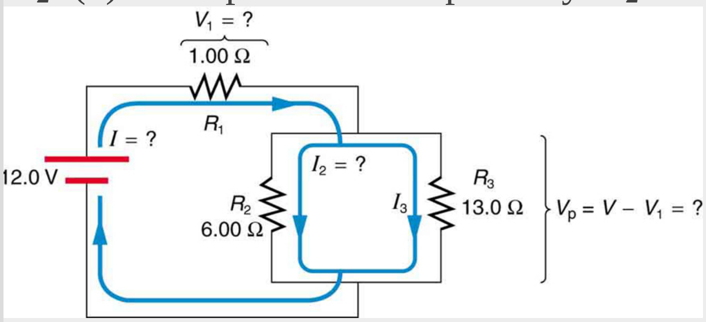


This combination of seven resistors has both series and parallel parts. Each is identified and reduced to an equivalent resistance, and these are further reduced until a single equivalent resistance is reached.

The simplest combination of series and parallel resistance, shown in [link], is also the most instructive, since it is found in many applications. For example, R_1 could be the resistance of wires from a car battery to its electrical devices, which are in parallel. R_2 and R_3 could be the starter motor and a passenger compartment light. We have previously assumed that wire resistance is negligible, but, when it is not, it has important effects, as the next example indicates.

Example:
Calculating Resistance, IR Drop, Current, and Power Dissipation: Combining Series and Parallel Circuits
[link] shows the resistors from the previous two examples wired in a different way—a combination of series and parallel. We can consider R_1 to be the resistance of wires leading to R_2 and R_3. (a) Find the total

resistance. (b) What is the IR drop in R_1? (c) Find the current I_2 through R_2. (d) What power is dissipated by R_2?



These three resistors are connected to a voltage source so that R_2 and R_3 are in parallel with one another and that combination is in series with R_1.

Strategy and Solution for (a)

To find the total resistance, we note that R_2 and R_3 are in parallel and their combination R_p is in series with R_1. Thus the total (equivalent) resistance of this combination is

Equation:

$$R_\mathrm{tot} = R_1 + R_\mathrm{p}.$$

First, we find R_p using the equation for resistors in parallel and entering known values:

Equation:

$$\frac{1}{R_\mathrm{p}} = \frac{1}{R_2} + \frac{1}{R_3} = \frac{1}{6.00\ \Omega} + \frac{1}{13.0\ \Omega} = \frac{0.2436}{\Omega}.$$

Inverting gives

Equation:

$$R_\mathrm{p} = \frac{1}{0.2436}\Omega = 4.11\ \Omega.$$

So the total resistance is

Equation:

$$R_{\text{tot}} = R_1 + R_{\text{p}} = 1.00\ \Omega + 4.11\ \Omega = 5.11\ \Omega.$$

Discussion for (a)
The total resistance of this combination is intermediate between the pure series and pure parallel values ($20.0\ \Omega$ and $0.804\ \Omega$, respectively) found for the same resistors in the two previous examples.
Strategy and Solution for (b)
To find the IR drop in R_1, we note that the full current I flows through R_1. Thus its IR drop is
Equation:

$$V_1 = \text{IR}_1.$$

We must find I before we can calculate V_1. The total current I is found using Ohm's law for the circuit. That is,
Equation:

$$I = \frac{V}{R_{\text{tot}}} = \frac{12.0\ \text{V}}{5.11\ \Omega} = 2.35\ \text{A}.$$

Entering this into the expression above, we get
Equation:

$$V_1 = \text{IR}_1 = (2.35\ \text{A})(1.00\ \Omega) = 2.35\ \text{V}.$$

Discussion for (b)
The voltage applied to R_2 and R_3 is less than the total voltage by an amount V_1. When wire resistance is large, it can significantly affect the operation of the devices represented by R_2 and R_3.
Strategy and Solution for (c)
To find the current through R_2, we must first find the voltage applied to it. We call this voltage V_{p}, because it is applied to a parallel combination of resistors. The voltage applied to both R_2 and R_3 is reduced by the amount V_1, and so it is
Equation:

$$V_p = V - V_1 = 12.0\ \text{V} - 2.35\ \text{V} = 9.65\ \text{V}.$$

Now the current I_2 through resistance R_2 is found using Ohm's law:
Equation:

$$I_2 = \frac{V_p}{R_2} = \frac{9.65\ \text{V}}{6.00\ \Omega} = 1.61\ \text{A}.$$

Discussion for (c)
The current is less than the 2.00 A that flowed through R_2 when it was connected in parallel to the battery in the previous parallel circuit example.
Strategy and Solution for (d)
The power dissipated by R_2 is given by
Equation:

$$P_2 = (I_2)^2 R_2 = (1.61\ \text{A})^2 (6.00\ \Omega) = 15.5\ \text{W}.$$

Discussion for (d)
The power is less than the 24.0 W this resistor dissipated when connected in parallel to the 12.0-V source.

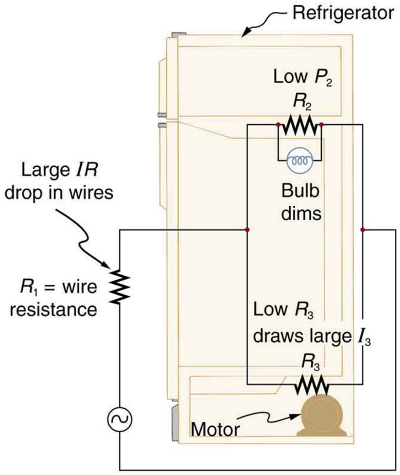
Practical Implications

One implication of this last example is that resistance in wires reduces the current and power delivered to a resistor. If wire resistance is relatively large, as in a worn (or a very long) extension cord, then this loss can be significant. If a large current is drawn, the IR drop in the wires can also be significant.

For example, when you are rummaging in the refrigerator and the motor comes on, the refrigerator light dims momentarily. Similarly, you can see the passenger compartment light dim when you start the engine of your car (although this may be due to resistance inside the battery itself).

What is happening in these high-current situations is illustrated in [link]. The device represented by R_3 has a very low resistance, and so when it is

switched on, a large current flows. This increased current causes a larger IR drop in the wires represented by R_1, reducing the voltage across the light bulb (which is R_2), which then dims noticeably.



Why do lights dim when a large appliance is switched on? The answer is that the large current the appliance motor draws causes a significant IR drop in the wires and reduces the voltage across the light.

Exercise:
Check Your Understanding

Problem:

Can any arbitrary combination of resistors be broken down into series and parallel combinations? See if you can draw a circuit diagram of resistors that cannot be broken down into combinations of series and parallel.

Solution:

No, there are many ways to connect resistors that are not combinations of series and parallel, including loops and junctions. In such cases Kirchhoff's rules, to be introduced in Kirchhoff's Rules, will allow you to analyze the circuit.

Note:
Problem-Solving Strategies for Series and Parallel Resistors

1. Draw a clear circuit diagram, labeling all resistors and voltage sources. This step includes a list of the knowns for the problem, since they are labeled in your circuit diagram.
2. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful.
3. Determine whether resistors are in series, parallel, or a combination of both series and parallel. Examine the circuit diagram to make this assessment. Resistors are in series if the same current must pass sequentially through them.
4. Use the appropriate list of major features for series or parallel connections to solve for the unknowns. There is one list for series and another for parallel. If your problem has a combination of series and parallel, reduce it in steps by considering individual groups of series or parallel connections, as done in this module and the examples. Special note: When finding R_{p}, the reciprocal must be taken with care.
5. Check to see whether the answers are reasonable and consistent. Units and numerical results must be reasonable. Total series resistance

should be greater, whereas total parallel resistance should be smaller, for example. Power should be greater for the same devices in parallel compared with series, and so on.

Section Summary

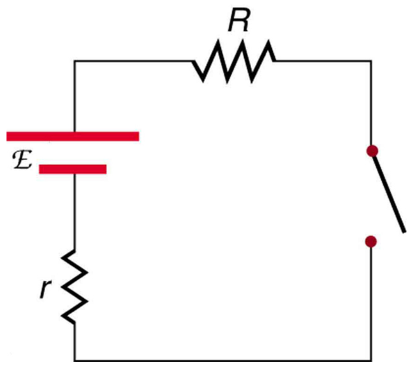
- The total resistance of an electrical circuit with resistors wired in a series is the sum of the individual resistances: $R_{\text{s}} = R_1 + R_2 + R_3 + \ldots.$
- Each resistor in a series circuit has the same amount of current flowing through it.
- The voltage drop, or power dissipation, across each individual resistor in a series is different, and their combined total adds up to the power source input.
- The total resistance of an electrical circuit with resistors wired in parallel is less than the lowest resistance of any of the components and can be determined using the formula:
  Equation:
  $$\frac{1}{R_{\text{p}}} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots.$$
- Each resistor in a parallel circuit has the same full voltage of the source applied to it.
- The current flowing through each resistor in a parallel circuit is different, depending on the resistance.
- If a more complex connection of resistors is a combination of series and parallel, it can be reduced to a single equivalent resistance by identifying its various parts as series or parallel, reducing each to its equivalent, and continuing until a single resistance is eventually reached.

Conceptual Questions

Exercise:

Problem:

A switch has a variable resistance that is nearly zero when closed and extremely large when open, and it is placed in series with the device it controls. Explain the effect the switch in [link] has on current when open and when closed.



A switch is ordinarily in series with a resistance and voltage source. Ideally, the switch has nearly zero resistance when closed but has an extremely large resistance when open. (Note that in this diagram, the script E represents the voltage (or electromotive force) of the battery.)

Exercise:

Problem: What is the voltage across the open switch in [link]?

Exercise:

Problem:

There is a voltage across an open switch, such as in [link]. Why, then, is the power dissipated by the open switch small?

Exercise:

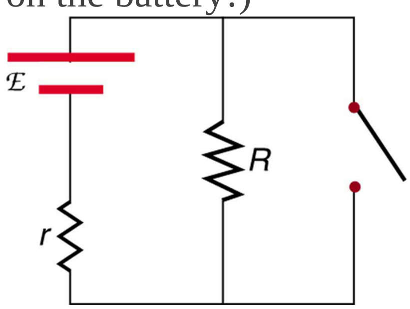
Problem:

Why is the power dissipated by a closed switch, such as in [link], small?

Exercise:

Problem:

A student in a physics lab mistakenly wired a light bulb, battery, and switch as shown in [link]. Explain why the bulb is on when the switch is open, and off when the switch is closed. (Do not try this—it is hard on the battery!)



A wiring mistake put this switch in parallel with the device represented by R. (Note that in this diagram, the script E represents the voltage (or

electromotive
force) of the
battery.)

Exercise:

Problem:

Knowing that the severity of a shock depends on the magnitude of the current through your body, would you prefer to be in series or parallel with a resistance, such as the heating element of a toaster, if shocked by it? Explain.

Exercise:

Problem:

Would your headlights dim when you start your car's engine if the wires in your automobile were superconductors? (Do not neglect the battery's internal resistance.) Explain.

Exercise:

Problem:

Some strings of holiday lights are wired in series to save wiring costs. An old version utilized bulbs that break the electrical connection, like an open switch, when they burn out. If one such bulb burns out, what happens to the others? If such a string operates on 120 V and has 40 identical bulbs, what is the normal operating voltage of each? Newer versions use bulbs that short circuit, like a closed switch, when they burn out. If one such bulb burns out, what happens to the others? If such a string operates on 120 V and has 39 remaining identical bulbs, what is then the operating voltage of each?

Exercise:

Problem:

If two household lightbulbs rated 60 W and 100 W are connected in series to household power, which will be brighter? Explain.

Exercise:

Problem:

Suppose you are doing a physics lab that asks you to put a resistor into a circuit, but all the resistors supplied have a larger resistance than the requested value. How would you connect the available resistances to attempt to get the smaller value asked for?

Exercise:

Problem:

Before World War II, some radios got power through a “resistance cord” that had a significant resistance. Such a resistance cord reduces the voltage to a desired level for the radio’s tubes and the like, and it saves the expense of a transformer. Explain why resistance cords become warm and waste energy when the radio is on.

Exercise:

Problem:

Some light bulbs have three power settings (not including zero), obtained from multiple filaments that are individually switched and wired in parallel. What is the minimum number of filaments needed for three power settings?

Problem Exercises

Note: Data taken from figures can be assumed to be accurate to three significant digits.

Exercise:

Problem:

(a) What is the resistance of ten $275\text{-}\Omega$ resistors connected in series? (b) In parallel?

Solution:

(a) $2.75\ \text{k}\Omega$

(b) $27.5\ \Omega$

Exercise:

Problem:

(a) What is the resistance of a 1.00×10^2 $\text{-}\Omega$, a $2.50\text{-k}\Omega$, and a $4.00\text{-k}\Omega$ resistor connected in series? (b) In parallel?

Exercise:

Problem:

What are the largest and smallest resistances you can obtain by connecting a $36.0\text{-}\Omega$, a $50.0\text{-}\Omega$, and a $700\text{-}\Omega$ resistor together?

Solution:

(a) $786\ \Omega$

(b) $20.3\ \Omega$

Exercise:

Problem:

An 1800-W toaster, a 1400-W electric frying pan, and a 75-W lamp are plugged into the same outlet in a 15-A, 120-V circuit. (The three devices are in parallel when plugged into the same socket.). (a) What current is drawn by each device? (b) Will this combination blow the 15-A fuse?

Exercise:

Problem:

Your car's 30.0-W headlight and 2.40-kW starter are ordinarily connected in parallel in a 12.0-V system. What power would one headlight and the starter consume if connected in series to a 12.0-V battery? (Neglect any other resistance in the circuit and any change in resistance in the two devices.)

Solution:

29.6 W

Exercise:

Problem:

(a) Given a 48.0-V battery and $24.0\text{-}\Omega$ and $96.0\text{-}\Omega$ resistors, find the current and power for each when connected in series. (b) Repeat when the resistances are in parallel.

Exercise:

Problem:

Referring to the example combining series and parallel circuits and [link], calculate I_3 in the following two different ways: (a) from the known values of I and I_2; (b) using Ohm's law for R_3. In both parts explicitly show how you follow the steps in the Problem-Solving Strategies for Series and Parallel Resistors.

Solution:

(a) 0.74 A

(b) 0.742 A

Exercise:

Problem:

Referring to [link]: (a) Calculate P_3 and note how it compares with P_3 found in the first two example problems in this module. (b) Find the total power supplied by the source and compare it with the sum of the powers dissipated by the resistors.

Exercise:

Problem:

Refer to [link] and the discussion of lights dimming when a heavy appliance comes on. (a) Given the voltage source is 120 V, the wire resistance is $0.400\ \Omega$, and the bulb is nominally 75.0 W, what power will the bulb dissipate if a total of 15.0 A passes through the wires when the motor comes on? Assume negligible change in bulb resistance. (b) What power is consumed by the motor?

Solution:

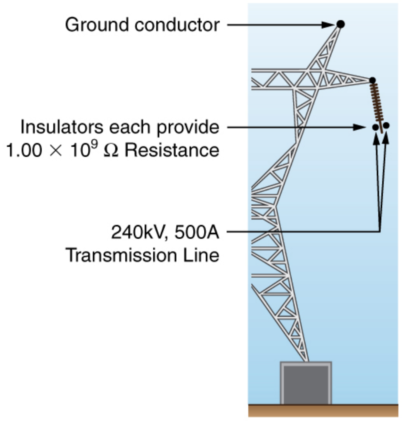
(a) 60.8 W

(b) 3.18 kW

Exercise:

Problem:

A 240-kV power transmission line carrying $5.00 \times 10^2\ \text{A}$ is hung from grounded metal towers by ceramic insulators, each having a $1.00{\times}10^9$ -Ω resistance. [link]. (a) What is the resistance to ground of 100 of these insulators? (b) Calculate the power dissipated by 100 of them. (c) What fraction of the power carried by the line is this? Explicitly show how you follow the steps in the Problem-Solving Strategies for Series and Parallel Resistors.



High-voltage (240-kV) transmission line carrying 5.00×10^2 A is hung from a grounded metal transmission tower. The row of ceramic insulators provide $1.00 \times 10^9 \Omega$ of resistance each.

Exercise:

Problem:

Show that if two resistors R_1 and R_2 are combined and one is much greater than the other ($R_1 >> R_2$): (a) Their series resistance is very nearly equal to the greater resistance R_1. (b) Their parallel resistance is very nearly equal to smaller resistance R_2.

Solution:

(a) $$\begin{array}{l} R_s = R_1 + R_2 \\ \Rightarrow R_s \approx R_1 (R_1 >> R_2) \end{array}$$

(b) $\frac{1}{R_{\text{p}}} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{R_1+R_2}{R_1 R_2}$,

so that

$$R_{\text{p}} = \frac{R_1 R_2}{R_1+R_2} \approx \frac{R_1 R_2}{R_1} = R_2 (R_1 >> R_2).$$

Exercise:

Problem: **Unreasonable Results**

Two resistors, one having a resistance of $145\ \Omega$, are connected in parallel to produce a total resistance of $150\ \Omega$. (a) What is the value of the second resistance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: **Unreasonable Results**

Two resistors, one having a resistance of $900\ \text{k}\Omega$, are connected in series to produce a total resistance of $0.500\ \text{M}\Omega$. (a) What is the value of the second resistance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) $-400\ \text{k}\Omega$

(b) Resistance cannot be negative.

(c) Series resistance is said to be less than one of the resistors, but it must be greater than any of the resistors.

Glossary

series

a sequence of resistors or other components wired into a circuit one after the other

resistor
: a component that provides resistance to the current flowing through an electrical circuit

resistance
: causing a loss of electrical power in a circuit

Ohm's law
: the relationship between current, voltage, and resistance within an electrical circuit: $V = \mathrm{IR}$

voltage
: the electrical potential energy per unit charge; electric pressure created by a power source, such as a battery

voltage drop
: the loss of electrical power as a current travels through a resistor, wire or other component

current
: the flow of charge through an electric circuit past a given point of measurement

Joule's law
: the relationship between potential electrical power, voltage, and resistance in an electrical circuit, given by: $P_e = \mathrm{IV}$

parallel
: the wiring of resistors or other components in an electrical circuit such that each component receives an equal voltage from the power source; often pictured in a ladder-shaped diagram, with each component on a rung of the ladder

Electromotive Force: Terminal Voltage

- Compare and contrast the voltage and the electromagnetic force of an electric power source.
- Describe what happens to the terminal voltage, current, and power delivered to a load as internal resistance of the voltage source increases (due to aging of batteries, for example).
- Explain why it is beneficial to use more than one voltage source connected in parallel.

When you forget to turn off your car lights, they slowly dim as the battery runs down. Why don't they simply blink off when the battery's energy is gone? Their gradual dimming implies that battery output voltage decreases as the battery is depleted.

Furthermore, if you connect an excessive number of 12-V lights in parallel to a car battery, they will be dim even when the battery is fresh and even if the wires to the lights have very low resistance. This implies that the battery's output voltage is reduced by the overload.

The reason for the decrease in output voltage for depleted or overloaded batteries is that all voltage sources have two fundamental parts—a source of electrical energy and an **internal resistance**. Let us examine both.

Electromotive Force

You can think of many different types of voltage sources. Batteries themselves come in many varieties. There are many types of mechanical/electrical generators, driven by many different energy sources, ranging from nuclear to wind. Solar cells create voltages directly from light, while thermoelectric devices create voltage from temperature differences.

A few voltage sources are shown in [link]. All such devices create a **potential difference** and can supply current if connected to a resistance. On the small scale, the potential difference creates an electric field that exerts force on charges, causing current. We thus use the name **electromotive force**, abbreviated emf.

Emf is not a force at all; it is a special type of potential difference. To be precise, the electromotive force (emf) is the potential difference of a source when no current is flowing. Units of emf are volts.

A variety of voltage sources (clockwise from top left): the Brazos Wind Farm in Fluvanna, Texas (credit: Leaflet, Wikimedia Commons); the Krasnoyarsk Dam in Russia (credit: Alex Polezhaev); a solar farm (credit: U.S. Department of Energy); and a group of nickel metal hydride batteries (credit: Tiaa Monto). The voltage output of each depends on its construction and load, and equals emf only if there is no load.

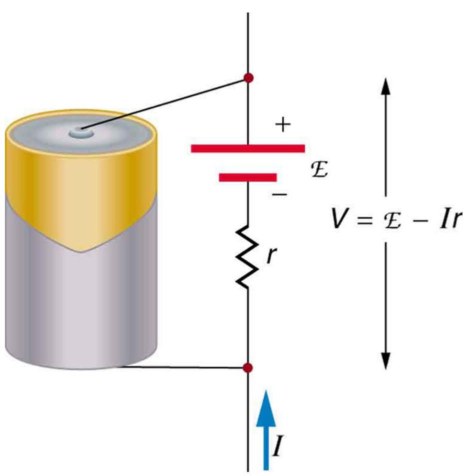
Electromotive force is directly related to the source of potential difference, such as the particular combination of chemicals in a battery. However, emf

differs from the voltage output of the device when current flows. The voltage across the terminals of a battery, for example, is less than the emf when the battery supplies current, and it declines further as the battery is depleted or loaded down. However, if the device's output voltage can be measured without drawing current, then output voltage will equal emf (even for a very depleted battery).

Internal Resistance

As noted before, a 12-V truck battery is physically larger, contains more charge and energy, and can deliver a larger current than a 12-V motorcycle battery. Both are lead-acid batteries with identical emf, but, because of its size, the truck battery has a smaller internal resistance r. Internal resistance is the inherent resistance to the flow of current within the source itself.

[link] is a schematic representation of the two fundamental parts of any voltage source. The emf (represented by a script E in the figure) and internal resistance r are in series. The smaller the internal resistance for a given emf, the more current and the more power the source can supply.



Any voltage source (in this case, a carbon-zinc dry cell) has an emf related to its source of

potential difference, and an internal resistance r related to its construction. (Note that the script E stands for emf.). Also shown are the output terminals across which the terminal voltage V is measured. Since $V = \text{emf} - \text{Ir}$, terminal voltage equals emf only if there is no current flowing.

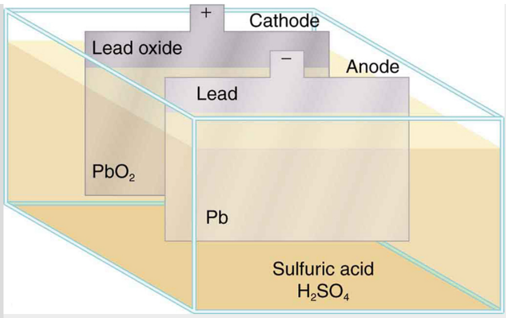
The internal resistance r can behave in complex ways. As noted, r increases as a battery is depleted. But internal resistance may also depend on the magnitude and direction of the current through a voltage source, its temperature, and even its history. The internal resistance of rechargeable nickel-cadmium cells, for example, depends on how many times and how deeply they have been depleted.

Note:

Things Great and Small: The Submicroscopic Origin of Battery Potential

Various types of batteries are available, with emfs determined by the combination of chemicals involved. We can view this as a molecular reaction (what much of chemistry is about) that separates charge.

The lead-acid battery used in cars and other vehicles is one of the most common types. A single cell (one of six) of this battery is seen in [link]. The cathode (positive) terminal of the cell is connected to a lead oxide plate, while the anode (negative) terminal is connected to a lead plate. Both plates are immersed in sulfuric acid, the electrolyte for the system.

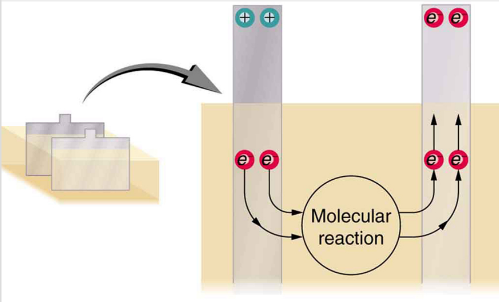


Artist's conception of a lead-acid cell. Chemical reactions in a lead-acid cell separate charge, sending negative charge to the anode, which is connected to the lead plates. The lead oxide plates are connected to the positive or cathode terminal of the cell. Sulfuric acid conducts the charge as well as participating in the chemical reaction.

The details of the chemical reaction are left to the reader to pursue in a chemistry text, but their results at the molecular level help explain the potential created by the battery. [link] shows the result of a single chemical reaction. Two electrons are placed on the anode, making it negative, provided that the cathode supplied two electrons. This leaves the cathode positively charged, because it has lost two electrons. In short, a separation of charge has been driven by a chemical reaction.

Note that the reaction will not take place unless there is a complete circuit to allow two electrons to be supplied to the cathode. Under many circumstances, these electrons come from the anode, flow through a resistance, and return to the cathode. Note also that since the chemical

reactions involve substances with resistance, it is not possible to create the emf without an internal resistance.



Artist’s conception of two electrons being forced onto the anode of a cell and two electrons being removed from the cathode of the cell. The chemical reaction in a lead-acid battery places two electrons on the anode and removes two from the cathode. It requires a closed circuit to proceed, since the two electrons must be supplied to the cathode.

Why are the chemicals able to produce a unique potential difference? Quantum mechanical descriptions of molecules, which take into account the types of atoms and numbers of electrons in them, are able to predict the energy states they can have and the energies of reactions between them.

In the case of a lead-acid battery, an energy of 2 eV is given to each electron sent to the anode. Voltage is defined as the electrical potential

energy divided by charge: $V = \frac{P_{\text{E}}}{q}$. An electron volt is the energy given to a single electron by a voltage of 1 V. So the voltage here is 2 V, since 2 eV is given to each electron. It is the energy produced in each molecular reaction that produces the voltage. A different reaction produces a different energy and, hence, a different voltage.

Terminal Voltage

The voltage output of a device is measured across its terminals and, thus, is called its **terminal voltage** V. Terminal voltage is given by

Equation:

$$V = \text{emf} - \text{Ir},$$

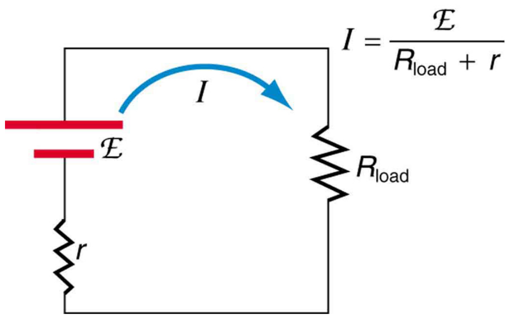
where r is the internal resistance and I is the current flowing at the time of the measurement.

I is positive if current flows away from the positive terminal, as shown in [link]. You can see that the larger the current, the smaller the terminal voltage. And it is likewise true that the larger the internal resistance, the smaller the terminal voltage.

Suppose a load resistance R_{load} is connected to a voltage source, as in [link]. Since the resistances are in series, the total resistance in the circuit is $R_{\text{load}} + r$. Thus the current is given by Ohm's law to be

Equation:

$$I = \frac{\text{emf}}{R_{\text{load}} + r}.$$



Schematic of a voltage source and its load R_{load}. Since the internal resistance r is in series with the load, it can significantly affect the terminal voltage and current delivered to the load. (Note that the script E stands for emf.)

We see from this expression that the smaller the internal resistance r, the greater the current the voltage source supplies to its load R_{load}. As batteries are depleted, r increases. If r becomes a significant fraction of the load resistance, then the current is significantly reduced, as the following example illustrates.

Example:
Calculating Terminal Voltage, Power Dissipation, Current, and Resistance: Terminal Voltage and Load
A certain battery has a 12.0-V emf and an internal resistance of $0.100\ \Omega$. (a) Calculate its terminal voltage when connected to a $10.0\text{-}\Omega$ load. (b) What is the terminal voltage when connected to a $0.500\text{-}\Omega$ load? (c) What power does the $0.500\text{-}\Omega$ load dissipate? (d) If the internal resistance grows

to 0.500 Ω, find the current, terminal voltage, and power dissipated by a 0.500-Ω load.

Strategy

The analysis above gave an expression for current when internal resistance is taken into account. Once the current is found, the terminal voltage can be calculated using the equation $V = \text{emf} - \text{Ir}$. Once current is found, the power dissipated by a resistor can also be found.

Solution for (a)

Entering the given values for the emf, load resistance, and internal resistance into the expression above yields

Equation:

$$I = \frac{\text{emf}}{R_{\text{load}} + r} = \frac{12.0\ \text{V}}{10.1\,\Omega} = 1.188\ \text{A}.$$

Enter the known values into the equation $V = \text{emf} - \text{Ir}$ to get the terminal voltage:

Equation:

$$\begin{array}{lll} V & = & \text{emf} - Ir = 12.0\ \text{V} - (1.188\ \text{A})(0.100\ \Omega) \\ & = & 11.9\ \text{V}. \end{array}$$

Discussion for (a)

The terminal voltage here is only slightly lower than the emf, implying that 10.0 Ω is a light load for this particular battery.

Solution for (b)

Similarly, with $R_{\text{load}} = 0.500\,\Omega$, the current is

Equation:

$$I = \frac{\text{emf}}{R_{\text{load}} + r} = \frac{12.0\ \text{V}}{0.600\,\Omega} = 20.0\ \text{A}.$$

The terminal voltage is now

Equation:

$$\begin{array}{lll} V & = & \text{emf} - Ir = 12.0\ \text{V} - (20.0\ \text{A})(0.100\ \Omega) \\ & = & 10.0\ \text{V}. \end{array}$$

Discussion for (b)

This terminal voltage exhibits a more significant reduction compared with emf, implying $0.500\,\Omega$ is a heavy load for this battery.

Solution for (c)

The power dissipated by the 0.500 - Ω load can be found using the formula $P = I^2 R$. Entering the known values gives

Equation:

$$P_{\text{load}} = I^2 R_{\text{load}} = (20.0\text{ A})^2 (0.500\,\Omega) = 2.00 \times 10^2\text{ W}.$$

Discussion for (c)

Note that this power can also be obtained using the expressions $\frac{V^2}{R}$ or IV, where V is the terminal voltage (10.0 V in this case).

Solution for (d)

Here the internal resistance has increased, perhaps due to the depletion of the battery, to the point where it is as great as the load resistance. As before, we first find the current by entering the known values into the expression, yielding

Equation:

$$I = \frac{\text{emf}}{R_{\text{load}} + r} = \frac{12.0\text{ V}}{1.00\,\Omega} = 12.0\text{ A}.$$

Now the terminal voltage is

Equation:

$$\begin{aligned} V &= \text{emf} - Ir = 12.0\text{ V} - (12.0\text{ A})(0.500\,\Omega) \\ &= 6.00\text{ V}, \end{aligned}$$

and the power dissipated by the load is

Equation:

$$P_{\text{load}} = I^2 R_{\text{load}} = (12.0\text{ A})^2 (0.500\,\Omega) = 72.0\text{ W}.$$

Discussion for (d)

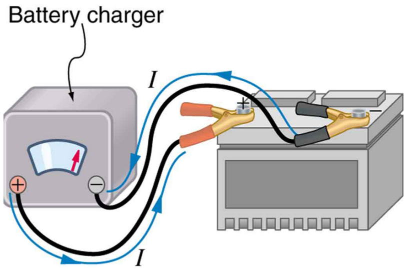
We see that the increased internal resistance has significantly decreased terminal voltage, current, and power delivered to a load.

Battery testers, such as those in [link], use small load resistors to intentionally draw current to determine whether the terminal voltage drops below an acceptable level. They really test the internal resistance of the battery. If internal resistance is high, the battery is weak, as evidenced by its low terminal voltage.

These two battery testers measure terminal voltage under a load to determine the condition of a battery. The large device is being used by a U.S. Navy electronics technician to test large batteries aboard the aircraft carrier USS *Nimitz* and has a small resistance that can dissipate large amounts of power. (credit: U.S. Navy photo by Photographer’s Mate Airman Jason A. Johnston) The small device is used on small batteries and has a digital display to indicate the acceptability of their terminal voltage. (credit: Keith Williamson)

Some batteries can be recharged by passing a current through them in the direction opposite to the current they supply to a resistance. This is done routinely in cars and batteries for small electrical appliances and electronic devices, and is represented pictorially in [link]. The voltage output of the battery charger must be greater than the emf of the battery to reverse current

through it. This will cause the terminal voltage of the battery to be greater than the emf, since $V = \text{emf} - \text{Ir}$, and I is now negative.



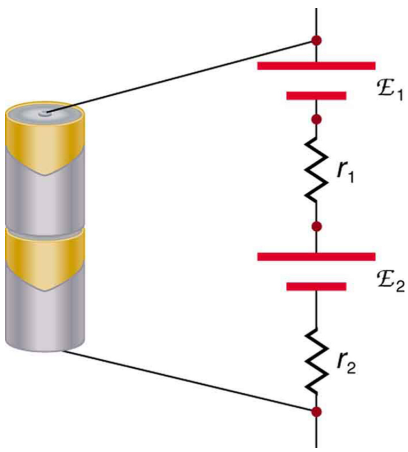
A car battery charger reverses the normal direction of current through a battery, reversing its chemical reaction and replenishing its chemical potential.

Multiple Voltage Sources

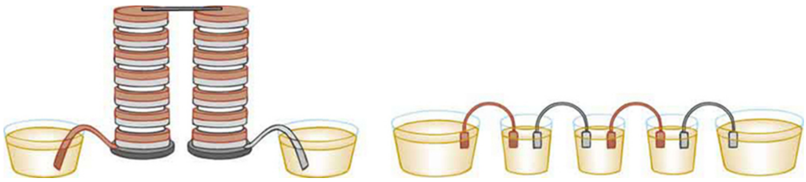
There are two voltage sources when a battery charger is used. Voltage sources connected in series are relatively simple. When voltage sources are in series, their internal resistances add and their emfs add algebraically. (See [link].) Series connections of voltage sources are common—for example, in flashlights, toys, and other appliances. Usually, the cells are in series in order to produce a larger total emf.

But if the cells oppose one another, such as when one is put into an appliance backward, the total emf is less, since it is the algebraic sum of the individual emfs.

A battery is a multiple connection of voltaic cells, as shown in [link]. The disadvantage of series connections of cells is that their internal resistances add. One of the authors once owned a 1957 MGA that had two 6-V batteries in series, rather than a single 12-V battery. This arrangement produced a large internal resistance that caused him many problems in starting the engine.

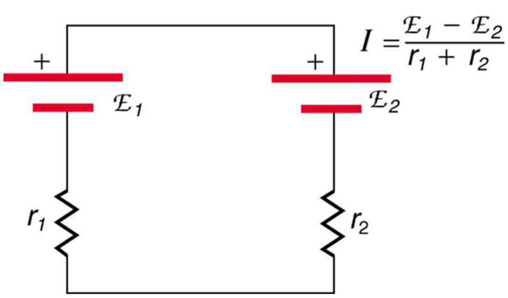


A series connection of two voltage sources. The emfs (each labeled with a script E) and internal resistances add, giving a total emf of $\text{emf}_1 + \text{emf}_2$ and a total internal resistance of $r_1 + r_2$.



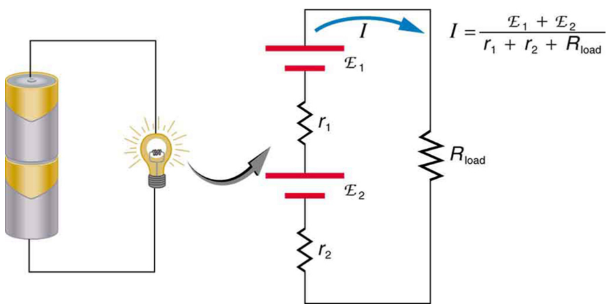
Batteries are multiple connections of individual cells, as shown in this modern rendition of an old print. Single cells, such as AA or C cells, are commonly called batteries, although this is technically incorrect.

If the *series* connection of two voltage sources is made into a complete circuit with the emfs in opposition, then a current of magnitude $I=\frac{(\text{emf}_1-\text{emf}_2)}{r_1+r_2}$ flows. See [link], for example, which shows a circuit exactly analogous to the battery charger discussed above. If two voltage sources in series with emfs in the same sense are connected to a load R_{load}, as in [link], then $I=\frac{(\text{emf}_1+\text{emf}_2)}{r_1+r_2+R_{\text{load}}}$ flows.



These two voltage sources are connected in series with their emfs in opposition. Current flows in the direction of the greater emf and is limited

to $I = \frac{(\text{emf}_1 - \text{emf}_2)}{r_1 + r_2}$ by the sum of the internal resistances. (Note that each emf is represented by script E in the figure.) A battery charger connected to a battery is an example of such a connection. The charger must have a larger emf than the battery to reverse current through it.

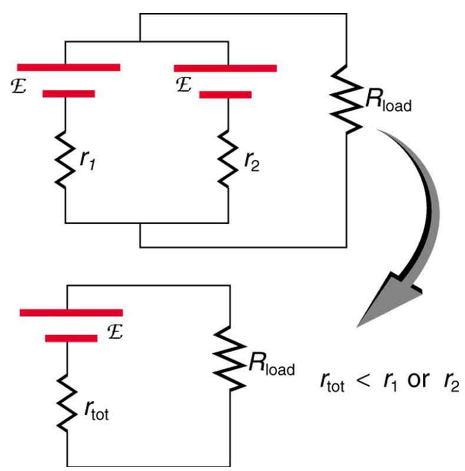


This schematic represents a flashlight with two cells (voltage sources) and a single bulb (load resistance) in series. The current that flows is $I = \frac{(\text{emf}_1 + \text{emf}_2)}{r_1 + r_2 + R_{\text{load}}}$. (Note that each emf is represented by script E in the figure.)

Note:
Take-Home Experiment: Flashlight Batteries
Find a flashlight that uses several batteries and find new and old batteries. Based on the discussions in this module, predict the brightness of the flashlight when different combinations of batteries are used. Do your predictions match what you observe? Now place new batteries in the flashlight and leave the flashlight switched on for several hours. Is the flashlight still quite bright? Do the same with the old batteries. Is the flashlight as bright when left on for the same length of time with old and new batteries? What does this say for the case when you are limited in the number of available new batteries?

[link] shows two voltage sources with identical emfs in parallel and connected to a load resistance. In this simple case, the total emf is the same as the individual emfs. But the total internal resistance is reduced, since the internal resistances are in parallel. The parallel connection thus can produce a larger current.

Here, $I = \frac{\text{emf}}{(r_{\text{tot}} + R_{\text{load}})}$ flows through the load, and r_{tot} is less than those of the individual batteries. For example, some diesel-powered cars use two 12-V batteries in parallel; they produce a total emf of 12 V but can deliver the larger current needed to start a diesel engine.



Two voltage sources with identical emfs (each labeled by script E) connected in parallel produce the same emf but have a smaller total internal resistance than the individual sources. Parallel combinations are often used to deliver more current. Here $I = \frac{\text{emf}}{(r_{\text{tot}} + R_{\text{load}})}$ flows through the load.

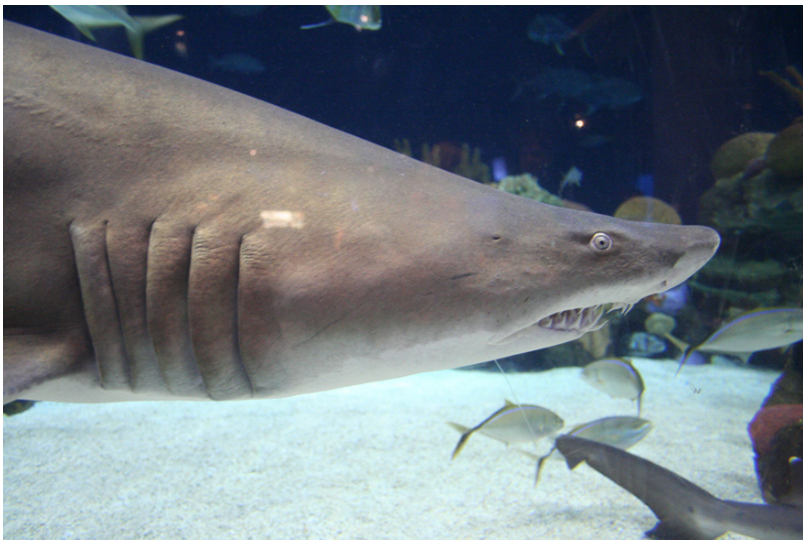
Animals as Electrical Detectors

A number of animals both produce and detect electrical signals. Fish, sharks, platypuses, and echidnas (spiny anteaters) all detect electric fields generated by nerve activity in prey. Electric eels produce their own emf through biological cells (electric organs) called electroplaques, which are arranged in both series and parallel as a set of batteries.

Electroplaques are flat, disk-like cells; those of the electric eel have a voltage of 0.15 V across each one. These cells are usually located toward the head or tail of the animal, although in the case of the electric eel, they are found along the entire body. The electroplaques in the South American eel are arranged in 140 rows, with each row stretching horizontally along the body and containing 5,000 electroplaques. This can yield an emf of approximately 600 V, and a current of 1 A—deadly.

The mechanism for detection of external electric fields is similar to that for producing nerve signals in the cell through depolarization and

repolarization—the movement of ions across the cell membrane. Within the fish, weak electric fields in the water produce a current in a gel-filled canal that runs from the skin to sensing cells, producing a nerve signal. The Australian platypus, one of the very few mammals that lay eggs, can detect fields of 30 $\frac{\text{mV}}{\text{m}}$, while sharks have been found to be able to sense a field in their snouts as small as 100 $\frac{\text{mV}}{\text{m}}$ ([link]). Electric eels use their own electric fields produced by the electroplaques to stun their prey or enemies.



Sand tiger sharks (*Carcharias taurus*), like this one at the Minnesota Zoo, use electroreceptors in their snouts to locate prey. (credit: Jim Winstead, Flickr)

Solar Cell Arrays

Another example dealing with multiple voltage sources is that of combinations of solar cells—wired in both series and parallel combinations to yield a desired voltage and current. Photovoltaic generation (PV), the conversion of sunlight directly into electricity, is based upon the

photoelectric effect, in which photons hitting the surface of a solar cell create an electric current in the cell.

Most solar cells are made from pure silicon—either as single-crystal silicon, or as a thin film of silicon deposited upon a glass or metal backing. Most single cells have a voltage output of about 0.5 V, while the current output is a function of the amount of sunlight upon the cell (the incident solar radiation—the insolation). Under bright noon sunlight, a current of about $100\ \text{mA/cm}^2$ of cell surface area is produced by typical single-crystal cells.

Individual solar cells are connected electrically in modules to meet electrical-energy needs. They can be wired together in series or in parallel—connected like the batteries discussed earlier. A solar-cell array or module usually consists of between 36 and 72 cells, with a power output of 50 W to 140 W.

The output of the solar cells is direct current. For most uses in a home, AC is required, so a device called an inverter must be used to convert the DC to AC. Any extra output can then be passed on to the outside electrical grid for sale to the utility.

Note:
Take-Home Experiment: Virtual Solar Cells
One can assemble a "virtual" solar cell array by using playing cards, or business or index cards, to represent a solar cell. Combinations of these cards in series and/or parallel can model the required array output. Assume each card has an output of 0.5 V and a current (under bright light) of 2 A. Using your cards, how would you arrange them to produce an output of 6 A at 3 V (18 W)?
Suppose you were told that you needed only 18 W (but no required voltage). Would you need more cards to make this arrangement?

Section Summary

- All voltage sources have two fundamental parts—a source of electrical energy that has a characteristic electromotive force (emf), and an internal resistance r.
- The emf is the potential difference of a source when no current is flowing.
- The numerical value of the emf depends on the source of potential difference.
- The internal resistance r of a voltage source affects the output voltage when a current flows.
- The voltage output of a device is called its terminal voltage V and is given by $V = \text{emf} - \text{Ir}$, where I is the electric current and is positive when flowing away from the positive terminal of the voltage source.
- When multiple voltage sources are in series, their internal resistances add and their emfs add algebraically.
- Solar cells can be wired in series or parallel to provide increased voltage or current, respectively.

Conceptual Questions

Exercise:

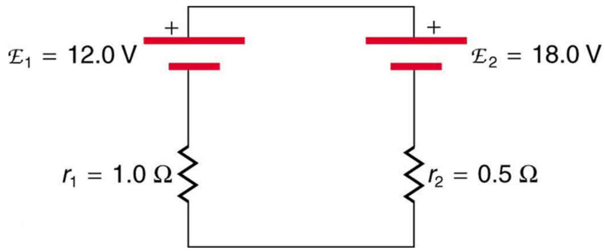
Problem:

Is every emf a potential difference? Is every potential difference an emf? Explain.

Exercise:

Problem:

Explain which battery is doing the charging and which is being charged in [link].



Exercise:

Problem:

Given a battery, an assortment of resistors, and a variety of voltage and current measuring devices, describe how you would determine the internal resistance of the battery.

Exercise:

Problem:

Two different 12-V automobile batteries on a store shelf are rated at 600 and 850 “cold cranking amps.” Which has the smallest internal resistance?

Exercise:

Problem:

What are the advantages and disadvantages of connecting batteries in series? In parallel?

Exercise:

Problem:

Semitractor trucks use four large 12-V batteries. The starter system requires 24 V, while normal operation of the truck’s other electrical components utilizes 12 V. How could the four batteries be connected to produce 24 V? To produce 12 V? Why is 24 V better than 12 V for starting the truck’s engine (a very heavy load)?

Problem Exercises

Exercise:

Problem:

Standard automobile batteries have six lead-acid cells in series, creating a total emf of 12.0 V. What is the emf of an individual lead-acid cell?

Solution:

2.00 V

Exercise:

Problem:

Carbon-zinc dry cells (sometimes referred to as non-alkaline cells) have an emf of 1.54 V, and they are produced as single cells or in various combinations to form other voltages. (a) How many 1.54-V cells are needed to make the common 9-V battery used in many small electronic devices? (b) What is the actual emf of the approximately 9-V battery? (c) Discuss how internal resistance in the series connection of cells will affect the terminal voltage of this approximately 9-V battery.

Exercise:

Problem:

What is the output voltage of a 3.0000-V lithium cell in a digital wristwatch that draws 0.300 mA, if the cell's internal resistance is $2.00\ \Omega$?

Solution:

2.9994 V

Exercise:

Problem:

(a) What is the terminal voltage of a large 1.54-V carbon-zinc dry cell used in a physics lab to supply 2.00 A to a circuit, if the cell's internal resistance is $0.100\ \Omega$? (b) How much electrical power does the cell produce? (c) What power goes to its load?

Exercise:

Problem:

What is the internal resistance of an automobile battery that has an emf of 12.0 V and a terminal voltage of 15.0 V while a current of 8.00 A is charging it?

Solution:

$0.375\ \Omega$

Exercise:

Problem:

(a) Find the terminal voltage of a 12.0-V motorcycle battery having a $0.600\text{-}\Omega$ internal resistance, if it is being charged by a current of 10.0 A. (b) What is the output voltage of the battery charger?

Exercise:

Problem:

A car battery with a 12-V emf and an internal resistance of $0.050\,\Omega$ is being charged with a current of 60 A. Note that in this process the battery is being charged. (a) What is the potential difference across its terminals? (b) At what rate is thermal energy being dissipated in the battery? (c) At what rate is electric energy being converted to chemical energy? (d) What are the answers to (a) and (b) when the battery is used to supply 60 A to the starter motor?

Exercise:

Problem:

The hot resistance of a flashlight bulb is $2.30\,\Omega$, and it is run by a 1.58-V alkaline cell having a 0.100-Ω internal resistance. (a) What current flows? (b) Calculate the power supplied to the bulb using I^2R_{bulb}. (c) Is this power the same as calculated using $\frac{V^2}{R_{\text{bulb}}}$?

Solution:

(a) 0.658 A

(b) 0.997 W

(c) 0.997 W; yes

Exercise:

Problem:

The label on a portable radio recommends the use of rechargeable nickel-cadmium cells (nicads), although they have a 1.25-V emf while alkaline cells have a 1.58-V emf. The radio has a 3.20-Ω resistance. (a) Draw a circuit diagram of the radio and its batteries. Now, calculate the power delivered to the radio. (b) When using Nicad cells each having an internal resistance of $0.0400\ \Omega$. (c) When using alkaline cells each having an internal resistance of $0.200\ \Omega$. (d) Does this difference seem significant, considering that the radio’s effective resistance is lowered when its volume is turned up?

Exercise:

Problem:

An automobile starter motor has an equivalent resistance of $0.0500\,\Omega$ and is supplied by a 12.0-V battery with a $0.0100\text{-}\Omega$ internal resistance. (a) What is the current to the motor? (b) What voltage is applied to it? (c) What power is supplied to the motor? (d) Repeat these calculations for when the battery connections are corroded and add $0.0900\,\Omega$ to the circuit. (Significant problems are caused by even small amounts of unwanted resistance in low-voltage, high-current applications.)

Solution:

(a) 200 A

(b) 10.0 V

(c) 2.00 kW

(d) $0.1000\ \Omega$; $80.0\ \text{A}$, $4.0\ \text{V}$, $320\ \text{W}$

Exercise:

Problem:

A child’s electronic toy is supplied by three 1.58-V alkaline cells having internal resistances of $0.0200\,\Omega$ in series with a 1.53-V carbon-zinc dry cell having a $0.100\text{-}\Omega$ internal resistance. The load resistance is $10.0\,\Omega$. (a) Draw a circuit diagram of the toy and its batteries. (b) What current flows? (c) How much power is supplied to the load? (d) What is the internal resistance of the dry cell if it goes bad, resulting in only 0.500 W being supplied to the load?

Exercise:

Problem:

(a) What is the internal resistance of a voltage source if its terminal voltage drops by 2.00 V when the current supplied increases by 5.00 A? (b) Can the emf of the voltage source be found with the information supplied?

Solution:

(a) $0.400\ \Omega$

(b) No, there is only one independent equation, so only r can be found.

Exercise:

Problem:

A person with body resistance between his hands of $10.0\ \text{k}\Omega$ accidentally grasps the terminals of a 20.0-kV power supply. (Do NOT do this!) (a) Draw a circuit diagram to represent the situation. (b) If the internal resistance of the power supply is $2000\ \Omega$, what is the current through his body? (c) What is the power dissipated in his body? (d) If the power supply is to be made safe by increasing its internal resistance, what should the internal resistance be for the maximum current in this situation to be 1.00 mA or less? (e) Will this modification compromise the effectiveness of the power supply for driving low-resistance devices? Explain your reasoning.

Exercise:

Problem:

Electric fish generate current with biological cells called electroplaques, which are physiological emf devices. The electroplaques in the South American eel are arranged in 140 rows, each row stretching horizontally along the body and each containing 5000 electroplaques. Each electroplaque has an emf of 0.15 V and internal resistance of $0.25\ \Omega$. If the water surrounding the fish has resistance of $800\ \Omega$, how much current can the eel produce in water from near its head to near its tail?

Exercise:

Problem: Integrated Concepts

A 12.0-V emf automobile battery has a terminal voltage of 16.0 V when being charged by a current of 10.0 A. (a) What is the battery’s internal resistance? (b) What power is dissipated inside the battery? (c) At what rate (in $^{\circ}\text{C/min}$) will its temperature increase if its mass is 20.0 kg and it has a specific heat of $0.300\ \text{kcal/kg}\cdot{}^{\circ}\text{C}$, assuming no heat escapes?

Exercise:

Problem: Unreasonable Results

A 1.58-V alkaline cell with a $0.200\text{-}\Omega$ internal resistance is supplying 8.50 A to a load. (a) What is its terminal voltage? (b) What is the value of the load resistance? (c) What is unreasonable about these results? (d) Which assumptions are unreasonable or inconsistent?

Solution:

(a) –0.120 V

(b) $-1.41\times10^{-2}\ \Omega$

(c) Negative terminal voltage; negative load resistance.

(d) The assumption that such a cell could provide 8.50 A is inconsistent with its internal resistance.

Exercise:

Problem: Unreasonable Results

(a) What is the internal resistance of a 1.54-V dry cell that supplies 1.00 W of power to a $15.0\text{-}\Omega$ bulb? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

electromotive force (emf)
: the potential difference of a source of electricity when no current is flowing; measured in volts

internal resistance
: the amount of resistance within the voltage source

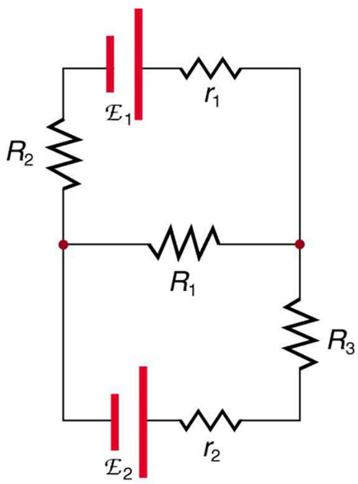
potential difference
: the difference in electric potential between two points in an electric circuit, measured in volts

terminal voltage
: the voltage measured across the terminals of a source of potential difference

Kirchhoff’s Rules

- Analyze a complex circuit using Kirchhoff’s rules, using the conventions for determining the correct signs of various terms.

Many complex circuits, such as the one in [link], cannot be analyzed with the series-parallel techniques developed in Resistors in Series and Parallel and Electromotive Force: Terminal Voltage. There are, however, two circuit analysis rules that can be used to analyze any circuit, simple or complex. These rules are special cases of the laws of conservation of charge and conservation of energy. The rules are known as **Kirchhoff’s rules**, after their inventor Gustav Kirchhoff (1824–1887).



This circuit cannot be reduced to a combination of series and parallel connections. Kirchhoff’s rules, special applications of the laws of conservation of charge and energy, can be

used to analyze it. (Note: The script E in the figure represents electromotive force, emf.)

Note:
Kirchhoff’s Rules

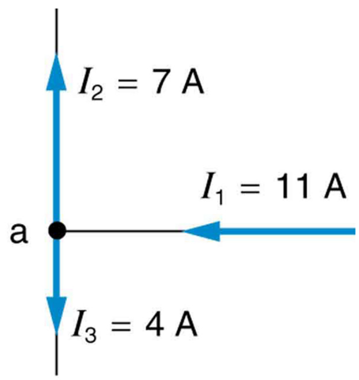
- Kirchhoff’s first rule—the junction rule. The sum of all currents entering a junction must equal the sum of all currents leaving the junction.
- Kirchhoff’s second rule—the loop rule. The algebraic sum of changes in potential around any closed circuit path (loop) must be zero.

Explanations of the two rules will now be given, followed by problem-solving hints for applying Kirchhoff’s rules, and a worked example that uses them.

Kirchhoff’s First Rule

Kirchhoff’s first rule (the **junction rule**) is an application of the conservation of charge to a junction; it is illustrated in [link]. Current is the flow of charge, and charge is conserved; thus, whatever charge flows into the junction must flow out. Kirchhoff’s first rule requires that $I_1 = I_2 + I_3$ (see figure). Equations like this can and will be used to analyze circuits and to solve circuit problems.

Note:
Making Connections: Conservation Laws
Kirchhoff’s rules for circuit analysis are applications of **conservation laws** to circuits. The first rule is the application of conservation of charge, while the second rule is the application of conservation of energy. Conservation laws, even used in a specific application, such as circuit analysis, are so basic as to form the foundation of that application.



$$I_1 = I_2 + I_3$$

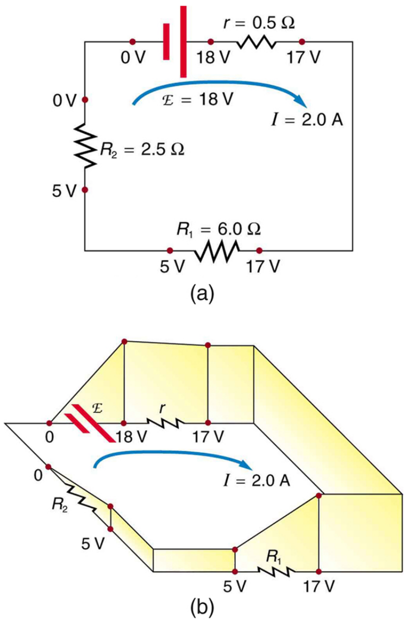
The junction rule. The diagram shows an example of Kirchhoff’s first rule where the sum of the currents into a junction equals the sum of the currents out of a junction. In this case, the current going into the junction splits and comes out as

two currents, so that $I_1 = I_2 + I_3$. Here I_1 must be 11 A, since I_2 is 7 A and I_3 is 4 A.

Kirchhoff's Second Rule

Kirchhoff's second rule (the **loop rule**) is an application of conservation of energy. The loop rule is stated in terms of potential, V, rather than potential energy, but the two are related since $\mathrm{PE_{elec}} = \mathrm{qV}$. Recall that **emf** is the potential difference of a source when no current is flowing. In a closed loop, whatever energy is supplied by emf must be transferred into other forms by devices in the loop, since there are no other ways in which energy can be transferred into or out of the circuit. [link] illustrates the changes in potential in a simple series circuit loop.

Kirchhoff's second rule requires $\mathrm{emf} - \mathrm{Ir} - \mathrm{IR_1} - \mathrm{IR_2} = 0$. Rearranged, this is $\mathrm{emf} = \mathrm{Ir} + IR_1 + IR_2$, which means the emf equals the sum of the IR (voltage) drops in the loop.



The loop rule. An example of Kirchhoff’s second rule where the sum of the changes in potential around a closed loop must be zero. (a) In this standard schematic of a simple series circuit, the emf supplies 18 V, which is reduced to zero by the resistances, with 1 V across the internal resistance, and 12 V and 5 V across the two load resistances, for a total of 18 V. (b) This perspective view represents the

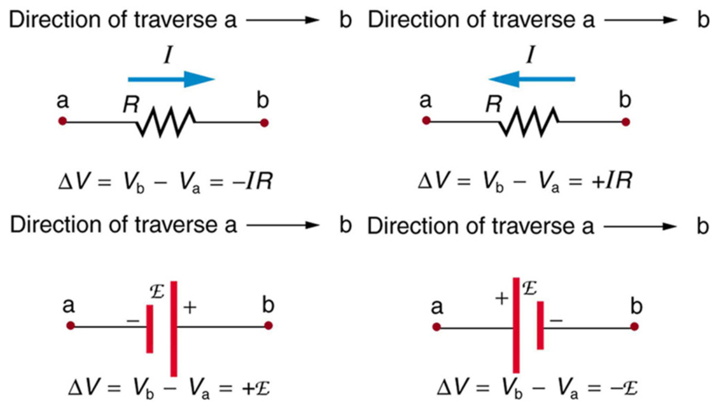
potential as something like a roller coaster, where charge is raised in potential by the emf and lowered by the resistances. (Note that the script E stands for emf.)

Applying Kirchhoff’s Rules

By applying Kirchhoff’s rules, we generate equations that allow us to find the unknowns in circuits. The unknowns may be currents, emfs, or resistances. Each time a rule is applied, an equation is produced. If there are as many independent equations as unknowns, then the problem can be solved. There are two decisions you must make when applying Kirchhoff’s rules. These decisions determine the signs of various quantities in the equations you obtain from applying the rules.

1. When applying Kirchhoff’s first rule, the junction rule, you must label the current in each branch and decide in what direction it is going. For example, in [link], [link], and [link], currents are labeled I_1, I_2, I_3, and I, and arrows indicate their directions. There is no risk here, for if you choose the wrong direction, the current will be of the correct magnitude but negative.
2. When applying Kirchhoff’s second rule, the loop rule, you must identify a closed loop and decide in which direction to go around it, clockwise or counterclockwise. For example, in [link] the loop was traversed in the same direction as the current (clockwise). Again, there is no risk; going around the circuit in the opposite direction reverses the sign of every term in the equation, which is like multiplying both sides of the equation by –1.

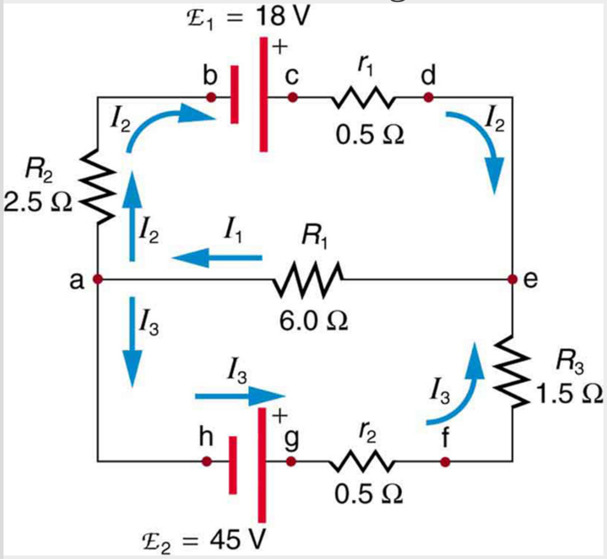
[link] and the following points will help you get the plus or minus signs right when applying the loop rule. Note that the resistors and emfs are traversed by going from a to b. In many circuits, it will be necessary to construct more than one loop. In traversing each loop, one needs to be consistent for the sign of the change in potential. (See [link].)



Each of these resistors and voltage sources is traversed from a to b. The potential changes are shown beneath each element and are explained in the text. (Note that the script E stands for emf.)

- When a resistor is traversed in the same direction as the current, the change in potential is $-IR$. (See [link].)
- When a resistor is traversed in the direction opposite to the current, the change in potential is $+IR$. (See [link].)
- When an emf is traversed from – to + (the same direction it moves positive charge), the change in potential is +emf. (See [link].)
- When an emf is traversed from + to – (opposite to the direction it moves positive charge), the change in potential is $-$emf. (See [link].)

Example:
Calculating Current: Using Kirchhoff’s Rules
Find the currents flowing in the circuit in [link].



This circuit is similar to that in [link], but the resistances and emfs are specified. (Each emf is denoted by script E.) The currents in each branch are labeled and assumed to move in the directions shown. This example uses Kirchhoff’s rules to find the currents.

Strategy
This circuit is sufficiently complex that the currents cannot be found using Ohm’s law and the series-parallel techniques—it is necessary to use Kirchhoff’s rules. Currents have been labeled I_1, I_2, and I_3 in the figure and assumptions have been made about their directions. Locations on the diagram have been labeled with letters a through h. In the solution we will apply the junction and loop rules, seeking three independent equations to allow us to solve for the three unknown currents.
Solution

We begin by applying Kirchhoff’s first or junction rule at point a. This gives

Equation:

$$I_1 = I_2 + I_3,$$

since I_1 flows into the junction, while I_2 and I_3 flow out. Applying the junction rule at e produces exactly the same equation, so that no new information is obtained. This is a single equation with three unknowns—three independent equations are needed, and so the loop rule must be applied.

Now we consider the loop abcdea. Going from a to b, we traverse R_2 in the same (assumed) direction of the current I_2, and so the change in potential is $-I_2R_2$. Then going from b to c, we go from – to +, so that the change in potential is $+\text{emf}_1$. Traversing the internal resistance r_1 from c to d gives $-I_2r_1$. Completing the loop by going from d to a again traverses a resistor in the same direction as its current, giving a change in potential of $-I_1R_1$.

The loop rule states that the changes in potential sum to zero. Thus,

Equation:

$$-I_2R_2 + \text{emf}_1 - I_2r_1 - I_1R_1 = -I_2(R_2 + r_1) + \text{emf}_1 - I_1R_1 = 0.$$

Substituting values from the circuit diagram for the resistances and emf, and canceling the ampere unit gives

Equation:

$$-3I_2 + 18 - 6I_1 = 0.$$

Now applying the loop rule to aefgha (we could have chosen abcdefgha as well) similarly gives

Equation:

$$+ I_1R_1 + I_3R_3 + I_3r_2 - \text{emf}_2 = +I_1R_1 + I_3(R_3 + r_2) - \text{emf}_2 = 0.$$

Note that the signs are reversed compared with the other loop, because elements are traversed in the opposite direction. With values entered, this becomes

Equation:

$$+ 6I_1 + 2I_3 - 45 = 0.$$

These three equations are sufficient to solve for the three unknown currents. First, solve the second equation for I_2:

Equation:

$$I_2 = 6 - 2I_1.$$

Now solve the third equation for I_3:

Equation:

$$I_3 = 22.5 - 3I_1.$$

Substituting these two new equations into the first one allows us to find a value for I_1:

Equation:

$$I_1 = I_2 + I_3 = (6 - 2I_1) + (22.5 - 3I_1) = 28.5 - 5I_1.$$

Combining terms gives

Equation:

$$6I_1 = 28.5, \text{ and}$$

Equation:

$$I_1 = 4.75 \text{ A}.$$

Substituting this value for I_1 back into the fourth equation gives

Equation:

$$I_2 = 6 - 2I_1 = 6 - 9.50$$

Equation:

$$I_2 = -3.50 \text{ A}.$$

The minus sign means I_2 flows in the direction opposite to that assumed in [link].

Finally, substituting the value for I_1 into the fifth equation gives

Equation:

$$I_3 = 22.5 - 3I_1 = 22.5 - 14.25$$

Equation:

$$I_3 = 8.25 \text{ A.}$$

Discussion
Just as a check, we note that indeed $I_1 = I_2 + I_3$. The results could also have been checked by entering all of the values into the equation for the abcdefgha loop.

Note:
Problem-Solving Strategies for Kirchhoff’s Rules

1. Make certain there is a clear circuit diagram on which you can label all known and unknown resistances, emfs, and currents. If a current is unknown, you must assign it a direction. This is necessary for determining the signs of potential changes. If you assign the direction incorrectly, the current will be found to have a negative value—no harm done.
2. Apply the junction rule to any junction in the circuit. Each time the junction rule is applied, you should get an equation with a current that does not appear in a previous application—if not, then the equation is redundant.
3. Apply the loop rule to as many loops as needed to solve for the unknowns in the problem. (There must be as many independent equations as unknowns.) To apply the loop rule, you must choose a direction to go around the loop. Then carefully and consistently determine the signs of the potential changes for each element using the four bulleted points discussed above in conjunction with [link].
4. Solve the simultaneous equations for the unknowns. This may involve many algebraic steps, requiring careful checking and rechecking.
5. Check to see whether the answers are reasonable and consistent. The numbers should be of the correct order of magnitude, neither

exceedingly large nor vanishingly small. The signs should be reasonable—for example, no resistance should be negative. Check to see that the values obtained satisfy the various equations obtained from applying the rules. The currents should satisfy the junction rule, for example.

The material in this section is correct in theory. We should be able to verify it by making measurements of current and voltage. In fact, some of the devices used to make such measurements are straightforward applications of the principles covered so far and are explored in the next modules. As we shall see, a very basic, even profound, fact results—making a measurement alters the quantity being measured.

Exercise:

Check Your Understanding

Problem:

Can Kirchhoff’s rules be applied to simple series and parallel circuits or are they restricted for use in more complicated circuits that are not combinations of series and parallel?

Solution:

Kirchhoff's rules can be applied to any circuit since they are applications to circuits of two conservation laws. Conservation laws are the most broadly applicable principles in physics. It is usually mathematically simpler to use the rules for series and parallel in simpler circuits so we emphasize Kirchhoff’s rules for use in more complicated situations. But the rules for series and parallel can be derived from Kirchhoff’s rules. Moreover, Kirchhoff’s rules can be expanded to devices other than resistors and emfs, such as capacitors, and are one of the basic analysis devices in circuit analysis.

Section Summary

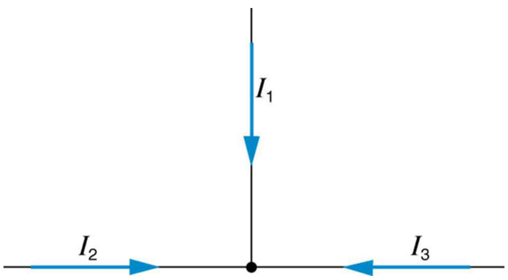
- Kirchhoff’s rules can be used to analyze any circuit, simple or complex.
- Kirchhoff’s first rule—the junction rule: The sum of all currents entering a junction must equal the sum of all currents leaving the junction.
- Kirchhoff’s second rule—the loop rule: The algebraic sum of changes in potential around any closed circuit path (loop) must be zero.
- The two rules are based, respectively, on the laws of conservation of charge and energy.
- When calculating potential and current using Kirchhoff’s rules, a set of conventions must be followed for determining the correct signs of various terms.
- The simpler series and parallel rules are special cases of Kirchhoff’s rules.

Conceptual Questions

Exercise:

Problem:

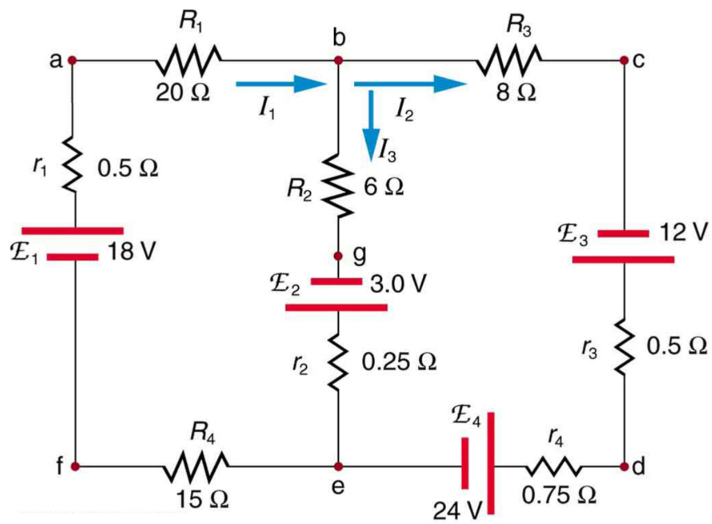
Can all of the currents going into the junction in [link] be positive? Explain.



Exercise:

Problem:

Apply the junction rule to junction b in [link]. Is any new information gained by applying the junction rule at e? (In the figure, each emf is represented by script E.)



Exercise:

Problem:

(a) What is the potential difference going from point a to point b in [link]? (b) What is the potential difference going from c to b? (c) From e to g? (d) From e to d?

Exercise:

Problem: Apply the loop rule to loop afedcba in [link].

Exercise:

Problem: Apply the loop rule to loops abgefa and cbgedc in [link].

Problem Exercises

Exercise:

Problem: Apply the loop rule to loop abcdefgha in [link].

Solution:
Equation:

$$-I_2R_2 + \text{emf}_1 - \text{I}_2r_1 + \text{I}_3R_3 + \text{I}_3r_2 - \text{emf}_2 = 0$$

Exercise:

Problem: Apply the loop rule to loop aedcba in [link].

Exercise:

Problem:

Verify the second equation in [link] by substituting the values found for the currents I_1 and I_2.

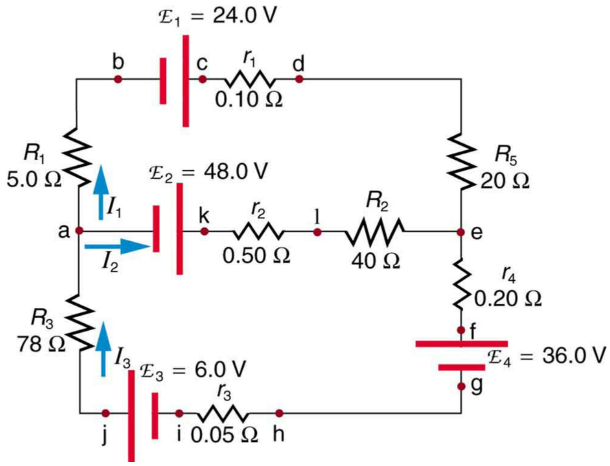
Exercise:

Problem:

Verify the third equation in [link] by substituting the values found for the currents I_1 and I_3.

Exercise:

Problem: Apply the junction rule at point a in [link].



Solution:
Equation:

$$I_3 = I_1 + I_2$$

Exercise:

Problem: Apply the loop rule to loop abcdefghija in [link].

Exercise:

Problem: Apply the loop rule to loop akledcba in [link].

Solution:
Equation:

$$\text{emf}_2 - I_2r_2 - I_2R_2 + I_1R_5 + I_1r_1 - \text{emf}_1 + I_1R_1 = 0$$

Exercise:

Problem:

Find the currents flowing in the circuit in [link]. Explicitly show how you follow the steps in the Problem-Solving Strategies for Series and Parallel Resistors.

Exercise:

Problem:

Solve [link], but use loop abcdefgha instead of loop akledcba. Explicitly show how you follow the steps in the Problem-Solving Strategies for Series and Parallel Resistors.

Solution:

(a) $I_1 = 4.75 \text{ A}$

(b) $I_2 = -3.5 \text{ A}$

(c) $I_3 = 8.25 \text{ A}$

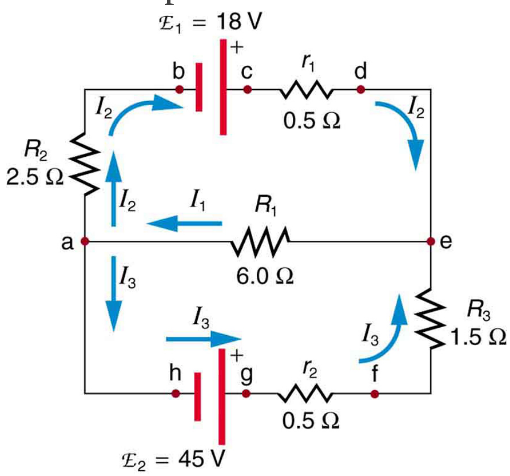
Exercise:

Problem: Find the currents flowing in the circuit in [link].

Exercise:

Problem: Unreasonable Results

Consider the circuit in [link], and suppose that the emfs are unknown and the currents are given to be $I_1 = 5.00 \text{ A}$, $I_2 = 3.0 \text{ A}$, and $I_3 = -2.00 \text{ A}$. (a) Could you find the emfs? (b) What is wrong with the assumptions?



Solution:

(a) No, you would get inconsistent equations to solve.

(b) $I_1 \neq I_2 + I_3$. The assumed currents violate the junction rule.

Glossary

Kirchhoff's rules
: a set of two rules, based on conservation of charge and energy, governing current and changes in potential in an electric circuit

junction rule
: Kirchhoff's first rule, which applies the conservation of charge to a junction; current is the flow of charge; thus, whatever charge flows into the junction must flow out; the rule can be stated $I_1 = I_2 + I_3$

loop rule
: Kirchhoff's second rule, which states that in a closed loop, whatever energy is supplied by emf must be transferred into other forms by devices in the loop, since there are no other ways in which energy can be transferred into or out of the circuit. Thus, the emf equals the sum of the IR (voltage) drops in the loop and can be stated: $\text{emf} = Ir + IR_1 + IR_2$

conservation laws
: require that energy and charge be conserved in a system

DC Voltmeters and Ammeters

- Explain why a voltmeter must be connected in parallel with the circuit.
- Draw a diagram showing an ammeter correctly connected in a circuit.
- Describe how a galvanometer can be used as either a voltmeter or an ammeter.
- Find the resistance that must be placed in series with a galvanometer to allow it to be used as a voltmeter with a given reading.
- Explain why measuring the voltage or current in a circuit can never be exact.

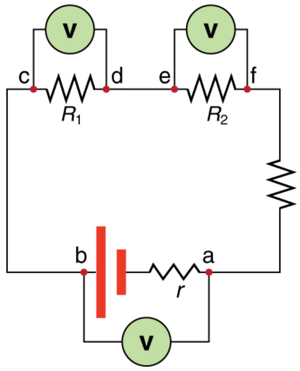
Voltmeters measure voltage, whereas **ammeters** measure current. Some of the meters in automobile dashboards, digital cameras, cell phones, and tuner-amplifiers are voltmeters or ammeters. (See [link].) The internal construction of the simplest of these meters and how they are connected to the system they monitor give further insight into applications of series and parallel connections.

The fuel and temperature gauges (far right and far left, respectively) in this 1996 Volkswagen are voltmeters that register the voltage output of "sender" units, which are hopefully proportional to the amount of gasoline in the tank and the engine

temperature. (credit: Christian Giersing)

Voltmeters are connected in parallel with whatever device’s voltage is to be measured. A parallel connection is used because objects in parallel experience the same potential difference. (See [link], where the voltmeter is represented by the symbol V.)

Ammeters are connected in series with whatever device’s current is to be measured. A series connection is used because objects in series have the same current passing through them. (See [link], where the ammeter is represented by the symbol A.)

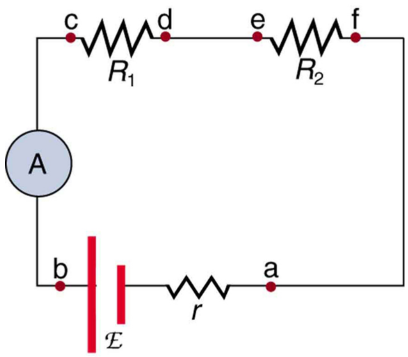


(a)

(b)

(a) To measure potential differences in this series circuit, the voltmeter (V) is placed in parallel with the

voltage source or either of the resistors. Note that terminal voltage is measured between points a and b. It is not possible to connect the voltmeter directly across the emf without including its internal resistance, r. (b) A digital voltmeter in use. (credit: Messtechniker, Wikimedia Commons)



An ammeter (A) is placed in series to measure current. All of the current in this circuit flows through the meter. The ammeter would have the same reading if located between points d and e or between points f and a as it does in the position shown. (Note that the script

capital E stands for emf, and r stands for the internal resistance of the source of potential difference.)

Analog Meters: Galvanometers

Analog meters have a needle that swivels to point at numbers on a scale, as opposed to **digital meters**, which have numerical readouts similar to a hand-held calculator. The heart of most analog meters is a device called a **galvanometer**, denoted by G. Current flow through a galvanometer, I_{G}, produces a proportional needle deflection. (This deflection is due to the force of a magnetic field upon a current-carrying wire.)

The two crucial characteristics of a given galvanometer are its resistance and current sensitivity. **Current sensitivity** is the current that gives a **full-scale deflection** of the galvanometer's needle, the maximum current that the instrument can measure. For example, a galvanometer with a current sensitivity of $50\ \mu\text{A}$ has a maximum deflection of its needle when $50\ \mu\text{A}$ flows through it, reads half-scale when $25\ \mu\text{A}$ flows through it, and so on.

If such a galvanometer has a $25\text{-}\Omega$ resistance, then a voltage of only $V=\text{IR}=(50\ \mu\text{A})(25\ \Omega)=1.25\ \text{mV}$ produces a full-scale reading. By connecting resistors to this galvanometer in different ways, you can use it as either a voltmeter or ammeter that can measure a broad range of voltages or currents.

Galvanometer as Voltmeter

[link] shows how a galvanometer can be used as a voltmeter by connecting it in series with a large resistance, R. The value of the resistance R is

determined by the maximum voltage to be measured. Suppose you want 10 V to produce a full-scale deflection of a voltmeter containing a $25\text{-}\Omega$ galvanometer with a $50\text{-}\mu\text{A}$ sensitivity. Then 10 V applied to the meter must produce a current of $50\ \mu\text{A}$. The total resistance must be

Equation:

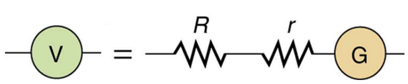
$$R_{\text{tot}} = R + r = \frac{V}{I} = \frac{10\text{ V}}{50\ \mu\text{A}} = 200\text{ k }\Omega\text{, or}$$

Equation:

$$R = R_{\text{tot}} - r = 200\text{ k}\Omega - 25\ \Omega \approx 200\text{ k}\Omega.$$

(R is so large that the galvanometer resistance, r, is nearly negligible.) Note that 5 V applied to this voltmeter produces a half-scale deflection by producing a $25\text{-}\mu\text{A}$ current through the meter, and so the voltmeter's reading is proportional to voltage as desired.

This voltmeter would not be useful for voltages less than about half a volt, because the meter deflection would be small and difficult to read accurately. For other voltage ranges, other resistances are placed in series with the galvanometer. Many meters have a choice of scales. That choice involves switching an appropriate resistance into series with the galvanometer.



A large resistance R placed in series with a galvanometer G produces a voltmeter, the full-scale deflection of which depends on the choice of R.

The larger the voltage to be measured, the larger R must be. (Note that r represents the internal resistance of the galvanometer.)

Galvanometer as Ammeter

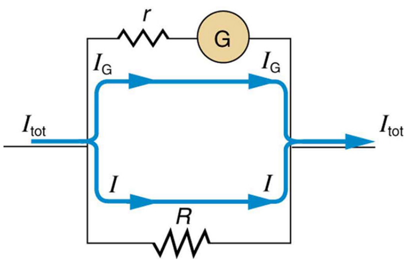
The same galvanometer can also be made into an ammeter by placing it in parallel with a small resistance R, often called the **shunt resistance**, as shown in [link]. Since the shunt resistance is small, most of the current passes through it, allowing an ammeter to measure currents much greater than those producing a full-scale deflection of the galvanometer.

Suppose, for example, an ammeter is needed that gives a full-scale deflection for 1.0 A, and contains the same $25\text{-}\Omega$ galvanometer with its $50\text{-}\mu\text{A}$ sensitivity. Since R and r are in parallel, the voltage across them is the same.

These IR drops are $\text{IR} = I_{\text{G}}r$ so that $\text{IR} = \frac{I_{\text{G}}}{I} = \frac{R}{r}$. Solving for R, and noting that I_{G} is 50 µA and I is 0.999950 A, we have

Equation:

$$R = r\frac{I_{\text{G}}}{I} = (25\,\Omega)\frac{50\ \mu\text{A}}{0.999950\ \text{A}} = 1.25\times10^{-3}\,\Omega.$$



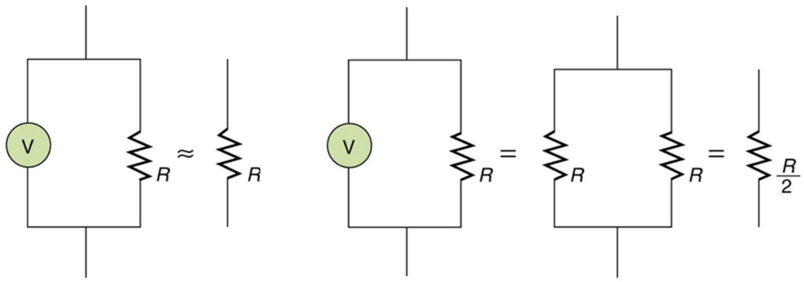
A small shunt resistance R placed in parallel with a galvanometer G produces an ammeter, the full-scale deflection of which depends on the choice of R. The larger the current to be measured, the smaller R must be. Most of the current (I) flowing through the meter is shunted through R to protect the galvanometer. (Note that r represents the internal resistance of the galvanometer.) Ammeters may also have multiple scales for greater flexibility in application. The various scales are

achieved by switching various shunt resistances in parallel with the galvanometer—the greater the maximum current to be measured, the smaller the shunt resistance must be.

Taking Measurements Alters the Circuit

When you use a voltmeter or ammeter, you are connecting another resistor to an existing circuit and, thus, altering the circuit. Ideally, voltmeters and ammeters do not appreciably affect the circuit, but it is instructive to examine the circumstances under which they do or do not interfere.

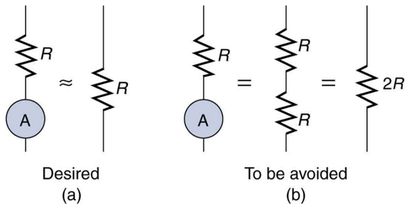
First, consider the voltmeter, which is always placed in parallel with the device being measured. Very little current flows through the voltmeter if its resistance is a few orders of magnitude greater than the device, and so the circuit is not appreciably affected. (See [link](a).) (A large resistance in parallel with a small one has a combined resistance essentially equal to the small one.) If, however, the voltmeter's resistance is comparable to that of the device being measured, then the two in parallel have a smaller resistance, appreciably affecting the circuit. (See [link](b).) The voltage across the device is not the same as when the voltmeter is out of the circuit.



(a) A voltmeter having a resistance much larger than the device ($R_{\text{Voltmeter}} >> R$) with which it is in parallel produces a parallel resistance essentially the same as the device and does not appreciably affect the circuit being measured. (b) Here the voltmeter has the same resistance as the device ($R_{\text{Voltmeter}} \cong R$), so that the parallel resistance is half of what it is when the voltmeter is not connected. This is an example of a significant alteration of the circuit and is to be avoided.

An ammeter is placed in series in the branch of the circuit being measured, so that its resistance adds to that branch. Normally, the ammeter's resistance is very small compared with the resistances of the devices in the circuit, and so the extra resistance is negligible. (See [link](a).) However, if very small load resistances are involved, or if the ammeter is not as low in resistance as it should be, then the total series resistance is significantly greater, and the current in the branch being measured is reduced. (See [link](b).)

A practical problem can occur if the ammeter is connected incorrectly. If it was put in parallel with the resistor to measure the current in it, you could possibly damage the meter; the low resistance of the ammeter would allow most of the current in the circuit to go through the galvanometer, and this current would be larger since the effective resistance is smaller.



(a) An ammeter normally has such a small resistance that the total series resistance in the branch being measured is not appreciably increased. The circuit is essentially unaltered compared with when the ammeter is absent. (b) Here the ammeter's resistance is the same as that of the branch, so that the total resistance is doubled and the current is half what it is without the ammeter. This significant alteration of the circuit is to be avoided.

One solution to the problem of voltmeters and ammeters interfering with the circuits being measured is to use galvanometers with greater sensitivity.

This allows construction of voltmeters with greater resistance and ammeters with smaller resistance than when less sensitive galvanometers are used.

There are practical limits to galvanometer sensitivity, but it is possible to get analog meters that make measurements accurate to a few percent. Note that the inaccuracy comes from altering the circuit, not from a fault in the meter.

Note:
Connections: Limits to Knowledge
Making a measurement alters the system being measured in a manner that produces uncertainty in the measurement. For macroscopic systems, such as the circuits discussed in this module, the alteration can usually be made negligibly small, but it cannot be eliminated entirely. For submicroscopic systems, such as atoms, nuclei, and smaller particles, measurement alters the system in a manner that cannot be made arbitrarily small. This actually limits knowledge of the system—even limiting what nature can know about itself. We shall see profound implications of this when the Heisenberg uncertainty principle is discussed in the modules on quantum mechanics.
There is another measurement technique based on drawing no current at all and, hence, not altering the circuit at all. These are called null measurements and are the topic of Null Measurements. Digital meters that employ solid-state electronics and null measurements can attain accuracies of one part in 10^6.

Exercise:
Check Your Understanding

Problem:

Digital meters are able to detect smaller currents than analog meters employing galvanometers. How does this explain their ability to measure voltage and current more accurately than analog meters?

Solution:

Since digital meters require less current than analog meters, they alter the circuit less than analog meters. Their resistance as a voltmeter can be far greater than an analog meter, and their resistance as an ammeter can be far less than an analog meter. Consult [link] and [link] and their discussion in the text.

Note:
PhET Explorations: Circuit Construction Kit (DC Only), Virtual Lab
Stimulate a neuron and monitor what happens. Pause, rewind, and move forward in time in order to observe the ions as they move across the neuron membrane.

Circuit Construction Kit (DC Only), Virtual Lab

Section Summary

- Voltmeters measure voltage, and ammeters measure current.
- A voltmeter is placed in parallel with the voltage source to receive full voltage and must have a large resistance to limit its effect on the circuit.
- An ammeter is placed in series to get the full current flowing through a branch and must have a small resistance to limit its effect on the circuit.

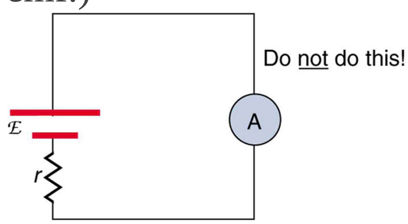
- Both can be based on the combination of a resistor and a galvanometer, a device that gives an analog reading of current.
- Standard voltmeters and ammeters alter the circuit being measured and are thus limited in accuracy.

Conceptual Questions

Exercise:

Problem:

Why should you not connect an ammeter directly across a voltage source as shown in [link]? (Note that script E in the figure stands for emf.)



Exercise:

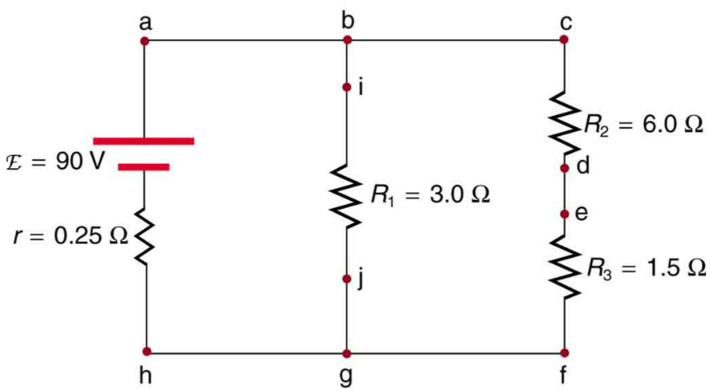
Problem:

Suppose you are using a multimeter (one designed to measure a range of voltages, currents, and resistances) to measure current in a circuit and you inadvertently leave it in a voltmeter mode. What effect will the meter have on the circuit? What would happen if you were measuring voltage but accidentally put the meter in the ammeter mode?

Exercise:

Problem:

Specify the points to which you could connect a voltmeter to measure the following potential differences in [link]: (a) the potential difference of the voltage source; (b) the potential difference across R_1; (c) across R_2; (d) across R_3; (e) across R_2 and R_3. Note that there may be more than one answer to each part.



Exercise:

Problem:

To measure currents in [link], you would replace a wire between two points with an ammeter. Specify the points between which you would place an ammeter to measure the following: (a) the total current; (b) the current flowing through R_1; (c) through R_2; (d) through R_3. Note that there may be more than one answer to each part.

Problem Exercises

Exercise:

Problem:

What is the sensitivity of the galvanometer (that is, what current gives a full-scale deflection) inside a voltmeter that has a $1.00\text{-M}\Omega$ resistance on its 30.0-V scale?

Solution:

$30\ \mu A$

Exercise:

Problem:

What is the sensitivity of the galvanometer (that is, what current gives a full-scale deflection) inside a voltmeter that has a $25.0\text{-k}\Omega$ resistance on its 100-V scale?

Exercise:

Problem:

Find the resistance that must be placed in series with a $25.0\text{-}\Omega$ galvanometer having a $50.0-\mu\text{A}$ sensitivity (the same as the one discussed in the text) to allow it to be used as a voltmeter with a 0.100-V full-scale reading.

Solution:

$1.98\ \text{k}\Omega$

Exercise:

Problem:

Find the resistance that must be placed in series with a $25.0\text{-}\Omega$ galvanometer having a $50.0\text{-}\mu\text{A}$ sensitivity (the same as the one discussed in the text) to allow it to be used as a voltmeter with a 3000-V full-scale reading. Include a circuit diagram with your solution.

Exercise:

Problem:

Find the resistance that must be placed in parallel with a $25.0\text{-}\Omega$ galvanometer having a $50.0\text{-}\mu\text{A}$ sensitivity (the same as the one discussed in the text) to allow it to be used as an ammeter with a 10.0-A full-scale reading. Include a circuit diagram with your solution.

Solution:
Equation:

$$1.25\times 10^{-4}\ \Omega$$

Exercise:

Problem:

Find the resistance that must be placed in parallel with a $25.0\text{-}\Omega$ galvanometer having a $50.0\text{-}\mu\text{A}$ sensitivity (the same as the one discussed in the text) to allow it to be used as an ammeter with a 300-mA full-scale reading.

Exercise:

Problem:

Find the resistance that must be placed in series with a $10.0\text{-}\Omega$ galvanometer having a $100\text{-}\mu\text{A}$ sensitivity to allow it to be used as a voltmeter with: (a) a 300-V full-scale reading, and (b) a 0.300-V full-scale reading.

Solution:

(a) $3.00\ \text{M}\Omega$

(b) $2.99\ \text{k}\Omega$

Exercise:

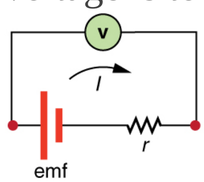
Problem:

Find the resistance that must be placed in parallel with a $10.0\text{-}\Omega$ galvanometer having a $100\text{-}\mu\text{A}$ sensitivity to allow it to be used as an ammeter with: (a) a 20.0-A full-scale reading, and (b) a 100-mA full-scale reading.

Exercise:

Problem:

Suppose you measure the terminal voltage of a 1.585-V alkaline cell having an internal resistance of $0.100\ \Omega$ by placing a $1.00\text{-k}\Omega$ voltmeter across its terminals. (See [link].) (a) What current flows? (b) Find the terminal voltage. (c) To see how close the measured terminal voltage is to the emf, calculate their ratio.



Solution:

(a) 1.58 mA

(b) 1.5848 V (need four digits to see the difference)

(c) 0.99990 (need five digits to see the difference from unity)

Exercise:

Problem:

Suppose you measure the terminal voltage of a 3.200-V lithium cell having an internal resistance of $5.00\ \Omega$ by placing a $1.00\text{-k}\Omega$ voltmeter across its terminals. (a) What current flows? (b) Find the terminal voltage. (c) To see how close the measured terminal voltage is to the emf, calculate their ratio.

Exercise:

Problem:

A certain ammeter has a resistance of $5.00 \times 10^{-5}\ \Omega$ on its 3.00-A scale and contains a $10.0\text{-}\Omega$ galvanometer. What is the sensitivity of the galvanometer?

Solution:

15.0 μA

Exercise:

Problem:

A 1.00-MΩ voltmeter is placed in parallel with a 75.0-kΩ resistor in a circuit. (a) Draw a circuit diagram of the connection. (b) What is the resistance of the combination? (c) If the voltage across the combination is kept the same as it was across the 75.0-kΩ resistor alone, what is the percent increase in current? (d) If the current through the combination is kept the same as it was through the 75.0-kΩ resistor alone, what is the percentage decrease in voltage? (e) Are the changes found in parts (c) and (d) significant? Discuss.

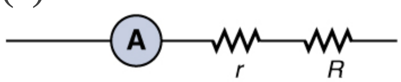
Exercise:

Problem:

A 0.0200-Ω ammeter is placed in series with a 10.00-Ω resistor in a circuit. (a) Draw a circuit diagram of the connection. (b) Calculate the resistance of the combination. (c) If the voltage is kept the same across the combination as it was through the 10.00-Ω resistor alone, what is the percent decrease in current? (d) If the current is kept the same through the combination as it was through the 10.00-Ω resistor alone, what is the percent increase in voltage? (e) Are the changes found in parts (c) and (d) significant? Discuss.

Solution:

(a)



(b) 10.02 Ω

(c) 0.9980, or a 2.0×10^{-1} percent decrease

(d) 1.002, or a 2.0×10^{-1} percent increase

(e) Not significant.

Exercise:

Problem: Unreasonable Results

Suppose you have a $40.0\text{-}\Omega$ galvanometer with a $25.0\text{-}\mu\text{A}$ sensitivity. (a) What resistance would you put in series with it to allow it to be used as a voltmeter that has a full-scale deflection for 0.500 mV? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Exercise:

Problem: Unreasonable Results

(a) What resistance would you put in parallel with a $40.0\text{-}\Omega$ galvanometer having a $25.0\text{-}\mu\text{A}$ sensitivity to allow it to be used as an ammeter that has a full-scale deflection for $10.0\text{-}\mu\text{A}$? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Solution:

(a) $-66.7\ \Omega$

(b) You can’t have negative resistance.

(c) It is unreasonable that I_{G} is greater than I_{tot} (see [link]). You cannot achieve a full-scale deflection using a current less than the sensitivity of the galvanometer.

Glossary

voltmeter
: an instrument that measures voltage

ammeter

an instrument that measures current

analog meter

a measuring instrument that gives a readout in the form of a needle movement over a marked gauge

digital meter

a measuring instrument that gives a readout in a digital form

galvanometer

an analog measuring device, denoted by G, that measures current flow using a needle deflection caused by a magnetic field force acting upon a current-carrying wire

current sensitivity

the maximum current that a galvanometer can read

full-scale deflection

the maximum deflection of a galvanometer needle, also known as current sensitivity; a galvanometer with a full-scale deflection of $50\ \mu A$ has a maximum deflection of its needle when $50\ \mu A$ flows through it

shunt resistance

a small resistance R placed in parallel with a galvanometer G to produce an ammeter; the larger the current to be measured, the smaller R must be; most of the current flowing through the meter is shunted through R to protect the galvanometer

Null Measurements

- Explain why a null measurement device is more accurate than a standard voltmeter or ammeter.
- Demonstrate how a Wheatstone bridge can be used to accurately calculate the resistance in a circuit.

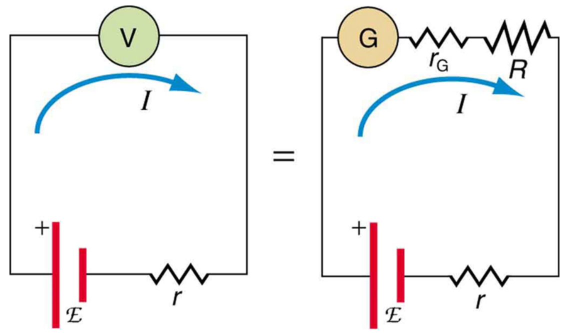
Standard measurements of voltage and current alter the circuit being measured, introducing uncertainties in the measurements. Voltmeters draw some extra current, whereas ammeters reduce current flow. **Null measurements** balance voltages so that there is no current flowing through the measuring device and, therefore, no alteration of the circuit being measured.

Null measurements are generally more accurate but are also more complex than the use of standard voltmeters and ammeters, and they still have limits to their precision. In this module, we shall consider a few specific types of null measurements, because they are common and interesting, and they further illuminate principles of electric circuits.

The Potentiometer

Suppose you wish to measure the emf of a battery. Consider what happens if you connect the battery directly to a standard voltmeter as shown in [link]. (Once we note the problems with this measurement, we will examine a null measurement that improves accuracy.) As discussed before, the actual quantity measured is the terminal voltage V, which is related to the emf of the battery by $V=\text{emf}-\text{Ir}$, where I is the current that flows and r is the internal resistance of the battery.

The emf could be accurately calculated if r were very accurately known, but it is usually not. If the current I could be made zero, then $V=\text{emf}$, and so emf could be directly measured. However, standard voltmeters need a current to operate; thus, another technique is needed.



An analog voltmeter attached to a battery draws a small but nonzero current and measures a terminal voltage that differs from the emf of the battery. (Note that the script capital E symbolizes electromotive force, or emf.) Since the internal resistance of the battery is not known precisely, it is not possible to calculate the emf precisely.

A **potentiometer** is a null measurement device for measuring potentials (voltages). (See [link].) A voltage source is connected to a resistor R, say, a long wire, and passes a constant current through it. There is a steady drop in potential (an IR drop) along the wire, so that a variable potential can be obtained by making contact at varying locations along the wire.

[link](b) shows an unknown emf_{x} (represented by script E_{x} in the figure) connected in series with a galvanometer. Note that emf_{x} opposes the other voltage source. The location of the contact point (see the arrow on the drawing) is adjusted until the galvanometer reads zero. When the galvanometer reads zero, $\text{emf}_{\text{x}} = IR_{\text{x}}$, where R_{x} is the resistance of the section of wire up to the contact point. Since no current flows through the

galvanometer, none flows through the unknown emf, and so emf_x is directly sensed.

Now, a very precisely known standard emf_s is substituted for emf_x, and the contact point is adjusted until the galvanometer again reads zero, so that $\text{emf}_\text{s} = IR_\text{s}$. In both cases, no current passes through the galvanometer, and so the current I through the long wire is the same. Upon taking the ratio $\frac{\text{emf}_\text{x}}{\text{emf}_\text{s}}$, I cancels, giving

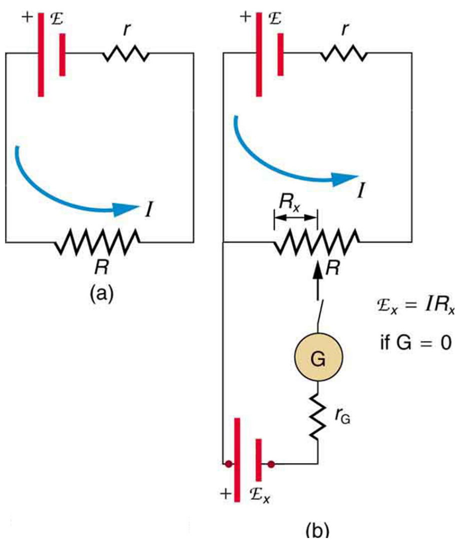
Equation:

$$\frac{\text{emf}_\text{x}}{\text{emf}_\text{s}} = \frac{IR_\text{x}}{IR_\text{s}} = \frac{R_\text{x}}{R_\text{s}}.$$

Solving for emf_x gives

Equation:

$$\text{emf}_\text{x} = \text{emf}_\text{s}\frac{R_\text{x}}{R_\text{s}}.$$



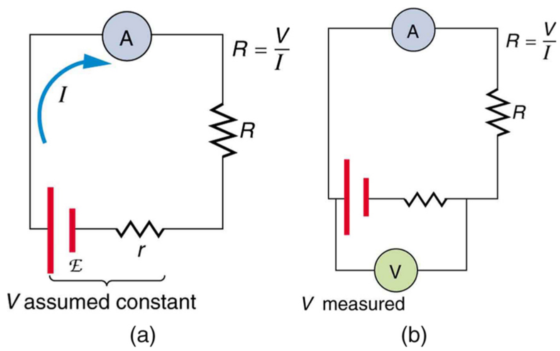
The potentiometer, a null measurement

device. (a) A voltage source connected to a long wire resistor passes a constant current I through it. (b) An unknown emf (labeled script E_x in the figure) is connected as shown, and the point of contact along R is adjusted until the galvanometer reads zero. The segment of wire has a resistance R_x and script $E_x = IR_x$, where I is unaffected by the connection since no current flows through the galvanometer. The unknown emf is thus proportional to the resistance of the wire segment.

Because a long uniform wire is used for R, the ratio of resistances R_x/R_s is the same as the ratio of the lengths of wire that zero the galvanometer for each emf. The three quantities on the right-hand side of the equation are now known or measured, and emf_x can be calculated. The uncertainty in this calculation can be considerably smaller than when using a voltmeter directly, but it is not zero. There is always some uncertainty in the ratio of resistances R_x/R_s and in the standard emf_s. Furthermore, it is not possible to tell when the galvanometer reads exactly zero, which introduces error into both R_x and R_s, and may also affect the current I.

Resistance Measurements and the Wheatstone Bridge

There is a variety of so-called **ohmmeters** that purport to measure resistance. What the most common ohmmeters actually do is to apply a voltage to a resistance, measure the current, and calculate the resistance using Ohm's law. Their readout is this calculated resistance. Two configurations for ohmmeters using standard voltmeters and ammeters are shown in [link]. Such configurations are limited in accuracy, because the meters alter both the voltage applied to the resistor and the current that flows through it.



Two methods for measuring resistance with standard meters. (a) Assuming a known voltage for the source, an ammeter measures current, and resistance is calculated as $R = \frac{V}{I}$. (b) Since the terminal voltage V varies with current, it is better to measure it. V is most accurately known when I is small, but I itself is most accurately known when it is large.

The **Wheatstone bridge** is a null measurement device for calculating resistance by balancing potential drops in a circuit. (See [link].) The device is called a bridge because the galvanometer forms a bridge between two branches. A variety of **bridge devices** are used to make null measurements in circuits.

Resistors R_1 and R_2 are precisely known, while the arrow through R_3 indicates that it is a variable resistance. The value of R_3 can be precisely read. With the unknown resistance R_{x} in the circuit, R_3 is adjusted until the galvanometer reads zero. The potential difference between points b and d is then zero, meaning that b and d are at the same potential. With no current running through the galvanometer, it has no effect on the rest of the circuit. So the branches abc and adc are in parallel, and each branch has the full voltage of the source. That is, the IR drops along abc and adc are the same. Since b and d are at the same potential, the IR drop along ad must equal the IR drop along ab. Thus,

Equation:

$$I_1R_1 = I_2R_3.$$

Again, since b and d are at the same potential, the IR drop along dc must equal the IR drop along bc. Thus,

Equation:

$$I_1R_2 = I_2R_{\mathrm{x}}.$$

Taking the ratio of these last two expressions gives

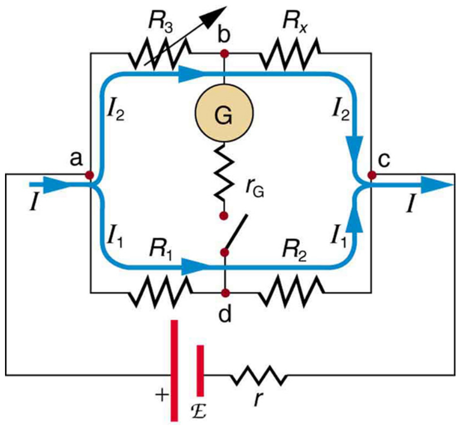
Equation:

$$\frac{I_1R_1}{I_1R_2} = \frac{I_2R_3}{I_2R_{\mathrm{x}}}.$$

Canceling the currents and solving for R_x yields

Equation:

$$R_x = R_3 \frac{R_2}{R_1}.$$



The Wheatstone bridge is used to calculate unknown resistances. The variable resistance R_3 is adjusted until the galvanometer reads zero with the switch closed. This simplifies the circuit, allowing R_x to be calculated based on the IR drops as discussed in the text.

This equation is used to calculate the unknown resistance when current through the galvanometer is zero. This method can be very accurate (often to four significant digits), but it is limited by two factors. First, it is not possible to get the current through the galvanometer to be exactly zero.

Second, there are always uncertainties in R_1, R_2, and R_3, which contribute to the uncertainty in R_x.

Exercise:

Check Your Understanding

Problem:

Identify other factors that might limit the accuracy of null measurements. Would the use of a digital device that is more sensitive than a galvanometer improve the accuracy of null measurements?

Solution:

One factor would be resistance in the wires and connections in a null measurement. These are impossible to make zero, and they can change over time. Another factor would be temperature variations in resistance, which can be reduced but not completely eliminated by choice of material. Digital devices sensitive to smaller currents than analog devices do improve the accuracy of null measurements because they allow you to get the current closer to zero.

Section Summary

- Null measurement techniques achieve greater accuracy by balancing a circuit so that no current flows through the measuring device.
- One such device, for determining voltage, is a potentiometer.
- Another null measurement device, for determining resistance, is the Wheatstone bridge.
- Other physical quantities can also be measured with null measurement techniques.

Conceptual questions

Exercise:

Problem:

Why can a null measurement be more accurate than one using standard voltmeters and ammeters? What factors limit the accuracy of null measurements?

Exercise:

Problem:

If a potentiometer is used to measure cell emfs on the order of a few volts, why is it most accurate for the standard emf_s to be the same order of magnitude and the resistances to be in the range of a few ohms?

Problem Exercises

Exercise:

Problem:

What is the emf_x of a cell being measured in a potentiometer, if the standard cell's emf is 12.0 V and the potentiometer balances for $R_\text{x} = 5.000\,\Omega$ and $R_\text{s} = 2.500\,\Omega$?

Solution:

24.0 V

Exercise:

Problem:

Calculate the emf_x of a dry cell for which a potentiometer is balanced when $R_\text{x} = 1.200\ \Omega$, while an alkaline standard cell with an emf of 1.600 V requires $R_\text{s} = 1.247\ \Omega$ to balance the potentiometer.

Exercise:

Problem:

When an unknown resistance R_{x} is placed in a Wheatstone bridge, it is possible to balance the bridge by adjusting R_3 to be $2500\ \Omega$. What is R_{x} if $\frac{R_2}{R_1} = 0.625$?

Solution:

$1.56\ \mathrm{k}\Omega$

Exercise:

Problem:

To what value must you adjust R_3 to balance a Wheatstone bridge, if the unknown resistance R_{x} is $100\,\Omega$, R_1 is $50.0\,\Omega$, and R_2 is $175\,\Omega$?

Exercise:

Problem:

(a) What is the unknown $\mathrm{emf}_{\mathrm{x}}$ in a potentiometer that balances when R_{x} is $10.0\,\Omega$, and balances when R_{s} is $15.0\,\Omega$ for a standard 3.000-V emf? (b) The same $\mathrm{emf}_{\mathrm{x}}$ is placed in the same potentiometer, which now balances when R_{s} is $15.0\,\Omega$ for a standard emf of 3.100 V. At what resistance R_{x} will the potentiometer balance?

Solution:

(a) 2.00 V

(b) $9.68\ \Omega$

Exercise:

Problem:

Suppose you want to measure resistances in the range from $10.0\ \Omega$ to $10.0\ \mathrm{k}\Omega$ using a Wheatstone bridge that has $\frac{R_2}{R_1} = 2.000$. Over what range should R_3 be adjustable?

Solution:
Equation:

$$\text{Range} = 5.00\ \Omega \text{ to } 5.00\ \text{k}\Omega$$

Glossary

null measurements
: methods of measuring current and voltage more accurately by balancing the circuit so that no current flows through the measurement device

potentiometer
: a null measurement device for measuring potentials (voltages)

ohmmeter
: an instrument that applies a voltage to a resistance, measures the current, calculates the resistance using Ohm’s law, and provides a readout of this calculated resistance

bridge device
: a device that forms a bridge between two branches of a circuit; some bridge devices are used to make null measurements in circuits

Wheatstone bridge
: a null measurement device for calculating resistance by balancing potential drops in a circuit

DC Circuits Containing Resistors and Capacitors

- Explain the importance of the time constant, τ , and calculate the time constant for a given resistance and capacitance.
- Explain why batteries in a flashlight gradually lose power and the light dims over time.
- Describe what happens to a graph of the voltage across a capacitor over time as it charges.
- Explain how a timing circuit works and list some applications.
- Calculate the necessary speed of a strobe flash needed to “stop” the movement of an object over a particular length.

When you use a flash camera, it takes a few seconds to charge the capacitor that powers the flash. The light flash discharges the capacitor in a tiny fraction of a second. Why does charging take longer than discharging? This question and a number of other phenomena that involve charging and discharging capacitors are discussed in this module.

RC Circuits

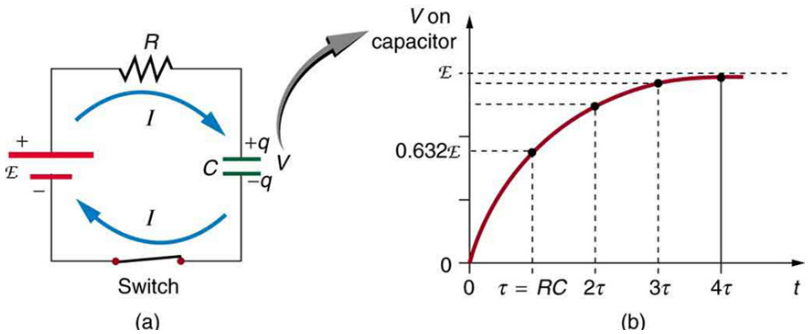
An $\mathbf{RC}$ **circuit** is one containing a **resistor** R and a **capacitor** C. The capacitor is an electrical component that stores electric charge.

[link] shows a simple RC circuit that employs a DC (direct current) voltage source. The capacitor is initially uncharged. As soon as the switch is closed, current flows to and from the initially uncharged capacitor. As charge increases on the capacitor plates, there is increasing opposition to the flow of charge by the repulsion of like charges on each plate.

In terms of voltage, this is because voltage across the capacitor is given by $V_{\text{c}} = Q/C$, where Q is the amount of charge stored on each plate and C is the **capacitance**. This voltage opposes the battery, growing from zero to the maximum emf when fully charged. The current thus decreases from its initial value of $I_0 = \frac{\text{emf}}{R}$ to zero as the voltage on the capacitor reaches the same value as the emf. When there is no current, there is no IR drop, and so the voltage on the capacitor must then equal the emf of the voltage source. This can also be explained with Kirchhoff’s second rule (the loop rule),

discussed in Kirchhoff's Rules, which says that the algebraic sum of changes in potential around any closed loop must be zero.

The initial current is $I_0 = \frac{\text{emf}}{R}$, because all of the IR drop is in the resistance. Therefore, the smaller the resistance, the faster a given capacitor will be charged. Note that the internal resistance of the voltage source is included in R, as are the resistances of the capacitor and the connecting wires. In the flash camera scenario above, when the batteries powering the camera begin to wear out, their internal resistance rises, reducing the current and lengthening the time it takes to get ready for the next flash.



(a) An RC circuit with an initially uncharged capacitor. Current flows in the direction shown (opposite of electron flow) as soon as the switch is closed. Mutual repulsion of like charges in the capacitor progressively slows the flow as the capacitor is charged, stopping the current when the capacitor is fully charged and $Q = C \cdot \text{emf}$. (b) A graph of voltage across the capacitor versus time, with the switch closing at time $t = 0$. (Note that in the two parts of the figure, the capital script E stands for emf, q stands for the charge stored on the capacitor, and τ is the RC time constant.)

Voltage on the capacitor is initially zero and rises rapidly at first, since the initial current is a maximum. [link](b) shows a graph of capacitor voltage versus time (t) starting when the switch is closed at $t=0$. The voltage approaches emf asymptotically, since the closer it gets to emf the less current flows. The equation for voltage versus time when charging a capacitor C through a resistor R, derived using calculus, is

Equation:

$$V=\text{emf}(1-e^{-t/\text{RC}})\text{ (charging)},$$

where V is the voltage across the capacitor, emf is equal to the emf of the DC voltage source, and the exponential e = 2.718 … is the base of the natural logarithm. Note that the units of RC are seconds. We define

Equation:

$$\tau=\text{RC},$$

where τ (the Greek letter tau) is called the time constant for an RC circuit. As noted before, a small resistance R allows the capacitor to charge faster. This is reasonable, since a larger current flows through a smaller resistance. It is also reasonable that the smaller the capacitor C, the less time needed to charge it. Both factors are contained in $\tau=\text{RC}$.

More quantitatively, consider what happens when $t=\tau=\text{RC}$. Then the voltage on the capacitor is

Equation:

$$V=\text{emf}\left(1-e^{-1}\right)=\text{emf}(1-0.368)=0.632\cdot\text{emf}.$$

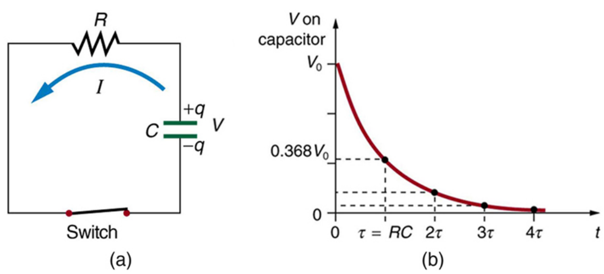
This means that in the time $\tau=\text{RC}$, the voltage rises to 0.632 of its final value. The voltage will rise 0.632 of the remainder in the next time τ. It is a characteristic of the exponential function that the final value is never reached, but 0.632 of the remainder to that value is achieved in every time, τ. In just a few multiples of the time constant τ, then, the final value is very nearly achieved, as the graph in [link](b) illustrates.

Discharging a Capacitor

Discharging a capacitor through a resistor proceeds in a similar fashion, as [link] illustrates. Initially, the current is $I_0 = \frac{V_0}{R}$, driven by the initial voltage V_0 on the capacitor. As the voltage decreases, the current and hence the rate of discharge decreases, implying another exponential formula for V. Using calculus, the voltage V on a capacitor C being discharged through a resistor R is found to be

Equation:

$$V = V_0 \, e^{-t/\text{RC}} (\text{discharging}).$$



(a) Closing the switch discharges the capacitor C through the resistor R. Mutual repulsion of like charges on each plate drives the current. (b) A graph of voltage across the capacitor versus time, with $V = V_0$ at $t = 0$. The voltage decreases exponentially, falling a fixed fraction of the way to zero in each subsequent time constant τ.

The graph in [link](b) is an example of this exponential decay. Again, the time constant is $\tau = \mathrm{RC}$. A small resistance R allows the capacitor to discharge in a small time, since the current is larger. Similarly, a small capacitance requires less time to discharge, since less charge is stored. In the first time interval $\tau = \mathrm{RC}$ after the switch is closed, the voltage falls to 0.368 of its initial value, since $V = V_0 \cdot e^{-1} = 0.368V_0$.

During each successive time τ, the voltage falls to 0.368 of its preceding value. In a few multiples of τ, the voltage becomes very close to zero, as indicated by the graph in [link](b).

Now we can explain why the flash camera in our scenario takes so much longer to charge than discharge; the resistance while charging is significantly greater than while discharging. The internal resistance of the battery accounts for most of the resistance while charging. As the battery ages, the increasing internal resistance makes the charging process even slower. (You may have noticed this.)

The flash discharge is through a low-resistance ionized gas in the flash tube and proceeds very rapidly. Flash photographs, such as in [link], can capture a brief instant of a rapid motion because the flash can be less than a microsecond in duration. Such flashes can be made extremely intense.

During World War II, nighttime reconnaissance photographs were made from the air with a single flash illuminating more than a square kilometer of enemy territory. The brevity of the flash eliminated blurring due to the surveillance aircraft’s motion. Today, an important use of intense flash lamps is to pump energy into a laser. The short intense flash can rapidly energize a laser and allow it to reemit the energy in another form.

This stop-motion photograph of a rufous hummingbird (*Selasphorus rufus*) feeding on a flower was obtained with an extremely brief and intense flash of light powered by the discharge of a capacitor through a gas. (credit: Dean E. Biggins, U.S. Fish and Wildlife Service)

Example:
Integrated Concept Problem: Calculating Capacitor Size—Strobe Lights

High-speed flash photography was pioneered by Doc Edgerton in the 1930s, while he was a professor of electrical engineering at MIT. You might have seen examples of his work in the amazing shots of hummingbirds in motion, a drop of milk splattering on a table, or a bullet penetrating an apple (see [link]). To stop the motion and capture these pictures, one needs a high-intensity, very short pulsed flash, as mentioned earlier in this module.

Suppose one wished to capture the picture of a bullet (moving at $5.0 \times 10^2 \text{ m/s}$) that was passing through an apple. The duration of the flash is related to the RC time constant, τ. What size capacitor would one

need in the RC circuit to succeed, if the resistance of the flash tube was $10.0\ \Omega$? Assume the apple is a sphere with a diameter of $8.0 \times 10^{-2}\ \mathrm{m}$.
Strategy
We begin by identifying the physical principles involved. This example deals with the strobe light, as discussed above. [link] shows the circuit for this probe. The characteristic time τ of the strobe is given as $\tau = \mathrm{RC}$.
Solution
We wish to find C, but we don't know τ. We want the flash to be on only while the bullet traverses the apple. So we need to use the kinematic equations that describe the relationship between distance x, velocity v, and time t:
Equation:

$$x = \mathrm{vt} \text{ or } t = \frac{x}{v}.$$

The bullet's velocity is given as $5.0 \times 10^{2}\ \mathrm{m/s}$, and the distance x is $8.0 \times 10^{-2}\ \mathrm{m}$. The traverse time, then, is
Equation:

$$t = \frac{x}{v} = \frac{8.0 \times 10^{-2}\ \mathrm{m}}{5.0 \times 10^{2}\ \mathrm{m/s}} = 1.6 \times 10^{-4}\ \mathrm{s}.$$

We set this value for the crossing time t equal to τ. Therefore,
Equation:

$$C = \frac{t}{R} = \frac{1.6{\times}10^{-4}\ \mathrm{s}}{10.0\ \Omega} = 16\ \mu\mathrm{F}.$$

(Note: Capacitance C is typically measured in farads, F, defined as Coulombs per volt. From the equation, we see that C can also be stated in units of seconds per ohm.)
Discussion
The flash interval of 160 μs (the traverse time of the bullet) is relatively easy to obtain today. Strobe lights have opened up new worlds from science to entertainment. The information from the picture of the apple and bullet was used in the Warren Commission Report on the assassination of

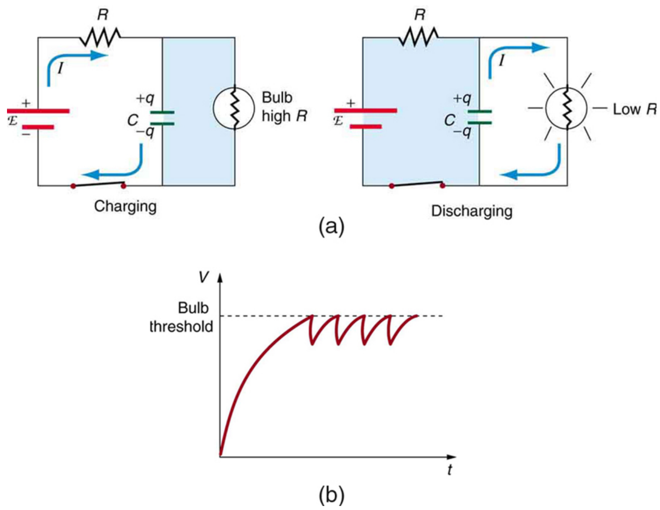
President John F. Kennedy in 1963 to confirm that only one bullet was fired.

RC Circuits for Timing

RC circuits are commonly used for timing purposes. A mundane example of this is found in the ubiquitous intermittent wiper systems of modern cars. The time between wipes is varied by adjusting the resistance in an RC circuit. Another example of an RC circuit is found in novelty jewelry, Halloween costumes, and various toys that have battery-powered flashing lights. (See [link] for a timing circuit.)

A more crucial use of RC circuits for timing purposes is in the artificial pacemaker, used to control heart rate. The heart rate is normally controlled by electrical signals generated by the sino-atrial (SA) node, which is on the wall of the right atrium chamber. This causes the muscles to contract and pump blood. Sometimes the heart rhythm is abnormal and the heartbeat is too high or too low.

The artificial pacemaker is inserted near the heart to provide electrical signals to the heart when needed with the appropriate time constant. Pacemakers have sensors that detect body motion and breathing to increase the heart rate during exercise to meet the body’s increased needs for blood and oxygen.



(a) The lamp in this RC circuit ordinarily has a very high resistance, so that the battery charges the capacitor as if the lamp were not there. When the voltage reaches a threshold value, a current flows through the lamp that dramatically reduces its resistance, and the capacitor discharges through the lamp as if the battery and charging resistor were not there. Once discharged, the process starts again, with the flash period determined by the RC constant τ. (b) A graph of voltage versus time for this circuit.

Example:
Calculating Time: *RC* Circuit in a Heart Defibrillator
A heart defibrillator is used to resuscitate an accident victim by discharging a capacitor through the trunk of her body. A simplified version of the

circuit is seen in [link]. (a) What is the time constant if an 8.00-μF capacitor is used and the path resistance through her body is $1.00 \times 10^3\ \Omega$? (b) If the initial voltage is 10.0 kV, how long does it take to decline to 5.00×10^2 V?

Strategy

Since the resistance and capacitance are given, it is straightforward to multiply them to give the time constant asked for in part (a). To find the time for the voltage to decline to 5.00×10^2 V, we repeatedly multiply the initial voltage by 0.368 until a voltage less than or equal to 5.00×10^2 V is obtained. Each multiplication corresponds to a time of τ seconds.

Solution for (a)

The time constant τ is given by the equation $\tau = \text{RC}$. Entering the given values for resistance and capacitance (and remembering that units for a farad can be expressed as s/Ω) gives

Equation:

$$\tau = \text{RC} = (1.00 \times 10^3\ \Omega)(8.00\ \mu F) = 8.00\ \text{ms.}$$

Solution for (b)

In the first 8.00 ms, the voltage (10.0 kV) declines to 0.368 of its initial value. That is:

Equation:

$$V = 0.368V_0 = 3.680 \times 10^3\ \text{V at}\ t = 8.00\ \text{ms.}$$

(Notice that we carry an extra digit for each intermediate calculation.) After another 8.00 ms, we multiply by 0.368 again, and the voltage is

Equation:

$$\begin{array}{rcl} V\prime & = & 0.368V \\ & = & (0.368)\left(3.680 \times 10^3\ \text{V}\right) \\ & = & 1.354 \times 10^3\ \text{V at}\ t = 16.0\ \text{ms.} \end{array}$$

Similarly, after another 8.00 ms, the voltage is

Equation:

$$\begin{array}{rcl} V'' & = & 0.368V' = (0.368)(1.354 \times 10^3 \text{ V}) \\ & = & 498 \text{ V at } t = 24.0 \text{ ms.} \end{array}$$

Discussion

So after only 24.0 ms, the voltage is down to 498 V, or 4.98% of its original value.Such brief times are useful in heart defibrillation, because the brief but intense current causes a brief but effective contraction of the heart. The actual circuit in a heart defibrillator is slightly more complex than the one in [link], to compensate for magnetic and AC effects that will be covered in Magnetism.

Exercise:
Check Your Understanding

Problem: When is the potential difference across a capacitor an emf?

Solution:

Only when the current being drawn from or put into the capacitor is zero. Capacitors, like batteries, have internal resistance, so their output voltage is not an emf unless current is zero. This is difficult to measure in practice so we refer to a capacitor’s voltage rather than its emf. But the source of potential difference in a capacitor is fundamental and it is an emf.

Note:

PhET Explorations: Circuit Construction Kit (DC only)

An electronics kit in your computer! Build circuits with resistors, light bulbs, batteries, and switches. Take measurements with the realistic ammeter and voltmeter. View the circuit as a schematic diagram, or switch to a life-like view.

https://archive.cnx.org/specials/f23ce496-c9d1-11e5-bdc8-bb04dc1eecb6/circuit-construction-kit-dc-only/#sim-cck

Section Summary

- An RC circuit is one that has both a resistor and a capacitor.
- The time constant τ for an RC circuit is $\tau = \mathrm{RC}$.
- When an initially uncharged ($V_0 = 0$ at $t = 0$) capacitor in series with a resistor is charged by a DC voltage source, the voltage rises, asymptotically approaching the emf of the voltage source; as a function of time,
  Equation:

$$V = \mathrm{emf}(1 - e^{-t/\mathrm{RC}})(\mathrm{charging}).$$

- Within the span of each time constant τ, the voltage rises by 0.632 of the remaining value, approaching the final voltage asymptotically.
- If a capacitor with an initial voltage V_0 is discharged through a resistor starting at $t = 0$, then its voltage decreases exponentially as given by
  Equation:

$$V = V_0 e^{-t/\mathrm{RC}}(\mathrm{discharging}).$$

- In each time constant τ, the voltage falls by 0.368 of its remaining initial value, approaching zero asymptotically.

Conceptual questions

Exercise:

Problem:

Regarding the units involved in the relationship $\tau = \mathrm{RC}$, verify that the units of resistance times capacitance are time, that is, $\Omega \cdot \mathrm{F} = \mathrm{s}$.

Exercise:

Problem:

The RC time constant in heart defibrillation is crucial to limiting the time the current flows. If the capacitance in the defibrillation unit is fixed, how would you manipulate resistance in the circuit to adjust the RC constant τ? Would an adjustment of the applied voltage also be needed to ensure that the current delivered has an appropriate value?

Exercise:

Problem:

When making an ECG measurement, it is important to measure voltage variations over small time intervals. The time is limited by the RC constant of the circuit—it is not possible to measure time variations shorter than RC. How would you manipulate R and C in the circuit to allow the necessary measurements?

Exercise:

Problem:

Draw two graphs of charge versus time on a capacitor. Draw one for charging an initially uncharged capacitor in series with a resistor, as in the circuit in [link], starting from $\mathrm{t} = 0$. Draw the other for discharging a capacitor through a resistor, as in the circuit in [link], starting at $\mathrm{t} = 0$, with an initial charge Q_0. Show at least two intervals of τ.

Exercise:

Problem:

When charging a capacitor, as discussed in conjunction with [link], how long does it take for the voltage on the capacitor to reach emf? Is this a problem?

Exercise:

Problem:

When discharging a capacitor, as discussed in conjunction with [link], how long does it take for the voltage on the capacitor to reach zero? Is this a problem?

Exercise:

Problem:

Referring to [link], draw a graph of potential difference across the resistor versus time, showing at least two intervals of τ. Also draw a graph of current versus time for this situation.

Exercise:

Problem:

A long, inexpensive extension cord is connected from inside the house to a refrigerator outside. The refrigerator doesn't run as it should. What might be the problem?

Exercise:

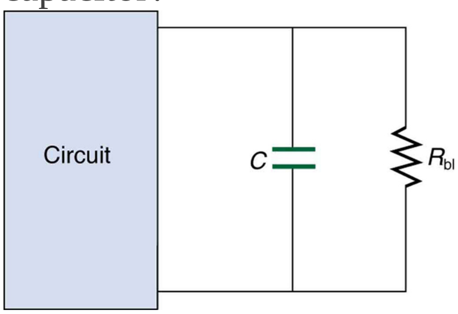
Problem:

In [link], does the graph indicate the time constant is shorter for discharging than for charging? Would you expect ionized gas to have low resistance? How would you adjust R to get a longer time between flashes? Would adjusting R affect the discharge time?

Exercise:

Problem:

An electronic apparatus may have large capacitors at high voltage in the power supply section, presenting a shock hazard even when the apparatus is switched off. A “bleeder resistor” is therefore placed across such a capacitor, as shown schematically in [link], to bleed the charge from it after the apparatus is off. Why must the bleeder resistance be much greater than the effective resistance of the rest of the circuit? How does this affect the time constant for discharging the capacitor?



A bleeder resistor R_{bl} discharges the capacitor in this electronic device once it is switched off.

Problem Exercises

Exercise:

Problem:

The timing device in an automobile’s intermittent wiper system is based on an RC time constant and utilizes a $0.500\text{-}\mu\mathrm{F}$ capacitor and a variable resistor. Over what range must R be made to vary to achieve time constants from 2.00 to 15.0 s?

Solution:

range $4.00 \text{ to } 30.0\ \mathrm{M\Omega}$

Exercise:

Problem:

A heart pacemaker fires 72 times a minute, each time a 25.0-nF capacitor is charged (by a battery in series with a resistor) to 0.632 of its full voltage. What is the value of the resistance?

Exercise:

Problem:

The duration of a photographic flash is related to an RC time constant, which is $0.100\ \mu s$ for a certain camera. (a) If the resistance of the flash lamp is $0.0400\,\Omega$ during discharge, what is the size of the capacitor supplying its energy? (b) What is the time constant for charging the capacitor, if the charging resistance is $800\ \mathrm{k\Omega}$?

Solution:

(a) $2.50\ \mu F$

(b) 2.00 s

Exercise:

Problem:

A 2.00- and a $7.50\text{-}\mu F$ capacitor can be connected in series or parallel, as can a 25.0- and a $100\text{-}\mathrm{k\Omega}$ resistor. Calculate the four RC time constants possible from connecting the resulting capacitance and resistance in series.

Exercise:

Problem:

After two time constants, what percentage of the final voltage, emf, is on an initially uncharged capacitor C, charged through a resistance R?

Solution:

86.5%

Exercise:

Problem:

A $500\text{-}\Omega$ resistor, an uncharged $1.50\text{-}\mu\text{F}$ capacitor, and a 6.16-V emf are connected in series. (a) What is the initial current? (b) What is the RC time constant? (c) What is the current after one time constant? (d) What is the voltage on the capacitor after one time constant?

Exercise:

Problem:

A heart defibrillator being used on a patient has an RC time constant of 10.0 ms due to the resistance of the patient and the capacitance of the defibrillator. (a) If the defibrillator has an $8.00\text{-}\mu\text{F}$ capacitance, what is the resistance of the path through the patient? (You may neglect the capacitance of the patient and the resistance of the defibrillator.) (b) If the initial voltage is 12.0 kV, how long does it take to decline to 6.00×10^2 V?

Solution:

(a) $1.25\ \text{k}\Omega$

(b) 30.0 ms

Exercise:

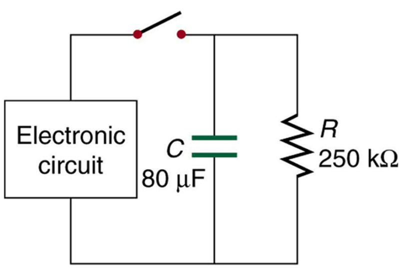
Problem:

An ECG monitor must have an RC time constant less than 1.00×10^2 μs to be able to measure variations in voltage over small time intervals. (a) If the resistance of the circuit (due mostly to that of the patient’s chest) is $1.00 \text{ k}\Omega$, what is the maximum capacitance of the circuit? (b) Would it be difficult in practice to limit the capacitance to less than the value found in (a)?

Exercise:

Problem:

[link] shows how a bleeder resistor is used to discharge a capacitor after an electronic device is shut off, allowing a person to work on the electronics with less risk of shock. (a) What is the time constant? (b) How long will it take to reduce the voltage on the capacitor to 0.250% (5% of 5%) of its full value once discharge begins? (c) If the capacitor is charged to a voltage V_0 through a $100\text{-}\Omega$ resistance, calculate the time it takes to rise to $0.865V_0$ (This is about two time constants.)



Solution:

(a) 20.0 s

(b) 120 s

(c) 16.0 ms

Exercise:

Problem:

Using the exact exponential treatment, find how much time is required to discharge a $250\text{-}\mu F$ capacitor through a $500\text{-}\Omega$ resistor down to 1.00% of its original voltage.

Exercise:

Problem:

Using the exact exponential treatment, find how much time is required to charge an initially uncharged 100-pF capacitor through a $75.0\text{-}M\Omega$ resistor to 90.0% of its final voltage.

Solution:

1.73×10^{-2} s

Exercise:

Problem: **Integrated Concepts**

If you wish to take a picture of a bullet traveling at 500 m/s, then a very brief flash of light produced by an RC discharge through a flash tube can limit blurring. Assuming 1.00 mm of motion during one RC constant is acceptable, and given that the flash is driven by a $600\text{-}\mu F$ capacitor, what is the resistance in the flash tube?

Solution:

3.33×10^{-3} Ω

Exercise:

Problem: **Integrated Concepts**

A flashing lamp in a Christmas earring is based on an RC discharge of a capacitor through its resistance. The effective duration of the flash is 0.250 s, during which it produces an average 0.500 W from an average

3.00 V. (a) What energy does it dissipate? (b) How much charge moves through the lamp? (c) Find the capacitance. (d) What is the resistance of the lamp?

Exercise:

Problem: Integrated Concepts

A $160\text{-}\mu\text{F}$ capacitor charged to 450 V is discharged through a $31.2\text{-k}\Omega$ resistor. (a) Find the time constant. (b) Calculate the temperature increase of the resistor, given that its mass is 2.50 g and its specific heat is $1.67\frac{\text{kJ}}{\text{kg}\cdot{}^{\text{o}}\text{C}}$, noting that most of the thermal energy is retained in the short time of the discharge. (c) Calculate the new resistance, assuming it is pure carbon. (d) Does this change in resistance seem significant?

Solution:

(a) 4.99 s

(b) 3.87^{o}C

(c) $31.1\text{ k}\Omega$

(d) No

Exercise:

Problem: Unreasonable Results

(a) Calculate the capacitance needed to get an RC time constant of 1.00×10^{3} s with a $0.100\text{-}\Omega$ resistor. (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Exercise:

Problem: Construct Your Own Problem

Consider a camera’s flash unit. Construct a problem in which you calculate the size of the capacitor that stores energy for the flash lamp. Among the things to be considered are the voltage applied to the capacitor, the energy needed in the flash and the associated charge needed on the capacitor, the resistance of the flash lamp during discharge, and the desired RC time constant.

Exercise:

Problem: Construct Your Own Problem

Consider a rechargeable lithium cell that is to be used to power a camcorder. Construct a problem in which you calculate the internal resistance of the cell during normal operation. Also, calculate the minimum voltage output of a battery charger to be used to recharge your lithium cell. Among the things to be considered are the emf and useful terminal voltage of a lithium cell and the current it should be able to supply to a camcorder.

Glossary

RC circuit
: a circuit that contains both a resistor and a capacitor

capacitor
: an electrical component used to store energy by separating electric charge on two opposing plates

capacitance
: the maximum amount of electric potential energy that can be stored (or separated) for a given electric potential

# Introduction to Magnetism
class="introduction"

The magnificent spectacle of the Aurora Borealis, or northern lights, glows in the northern sky above Bear Lake near Eielson Air Force Base, Alaska. Shaped by the Earth’s magnetic field, this light is produced by radiation spewed from solar storms. (credit: Senior Airman Joshua Strang, via Flickr)

One evening, an Alaskan sticks a note to his refrigerator with a small magnet. Through the kitchen window, the Aurora Borealis glows in the night sky. This grand spectacle is shaped by the same force that holds the note to the refrigerator.

People have been aware of magnets and magnetism for thousands of years. The earliest records date to well before the time of Christ, particularly in a region of Asia Minor called Magnesia (the name of this region is the source of words like *magnetic*). Magnetic rocks found in Magnesia, which is now part of western Turkey, stimulated interest during ancient times. A practical application for magnets was found later, when they were employed as navigational compasses. The use of magnets in compasses resulted not only in improved long-distance sailing, but also in the names of “north” and “south” being given to the two types of magnetic poles.

Today magnetism plays many important roles in our lives. Physicists’ understanding of magnetism has enabled the development of technologies that affect our everyday lives. The iPod in your purse or backpack, for example, wouldn’t have been possible without the applications of magnetism and electricity on a small scale.

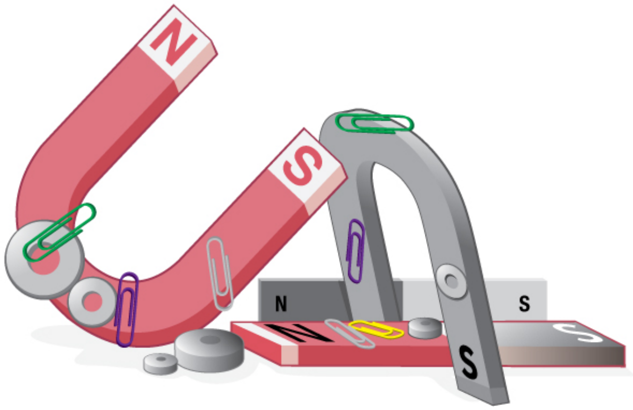
The discovery that weak changes in a magnetic field in a thin film of iron and chromium could bring about much larger changes in electrical resistance was one of the first large successes of nanotechnology. The 2007 Nobel Prize in Physics went to Albert Fert from France and Peter Grunberg from Germany for this discovery of *giant magnetoresistance* and its applications to computer memory.

All electric motors, with uses as diverse as powering refrigerators, starting cars, and moving elevators, contain magnets. Generators, whether producing hydroelectric power or running bicycle lights, use magnetic fields. Recycling facilities employ magnets to separate iron from other refuse. Hundreds of millions of dollars are spent annually on magnetic containment of fusion as a future energy source. Magnetic resonance imaging (MRI) has become an important diagnostic tool in the field of medicine, and the use of magnetism to explore brain activity is a subject of contemporary research and development. The list of applications also includes computer hard drives, tape recording, detection of inhaled asbestos, and levitation of high-speed trains. Magnetism is used to explain atomic energy levels, cosmic rays, and charged particles trapped in the Van Allen belts. Once again, we will find all these disparate phenomena are linked by a small number of underlying physical principles.

Engineering of technology like iPods would not be possible without a deep understanding magnetism. (credit: Jesse! S?, Flickr)

Magnets

- Describe the difference between the north and south poles of a magnet.
- Describe how magnetic poles interact with each other.

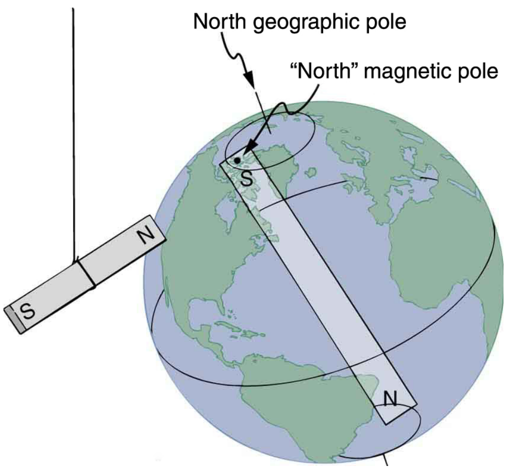


Magnets come in various shapes, sizes, and strengths. All have both a north pole and a south pole. There is never an isolated pole (a monopole).

All magnets attract iron, such as that in a refrigerator door. However, magnets may attract or repel other magnets. Experimentation shows that all magnets have two poles. If freely suspended, one pole will point toward the north. The two poles are thus named the **north magnetic pole** and the **south magnetic pole** (or more properly, north-seeking and south-seeking poles, for the attractions in those directions).

Note:
Universal Characteristics of Magnets and Magnetic Poles
It is a universal characteristic of all magnets that *like poles repel and unlike poles attract*. (Note the similarity with electrostatics: unlike charges attract and like charges repel.)

Further experimentation shows that it is *impossible to separate north and south poles* in the manner that + and − charges can be separated.



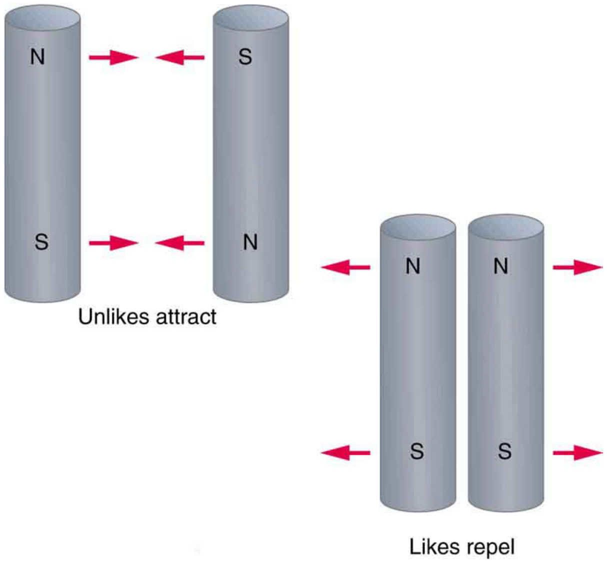
One end of a bar magnet is suspended from a thread that points toward north. The magnet's two poles are labeled N and S for north-seeking and south-seeking poles, respectively.

Note:

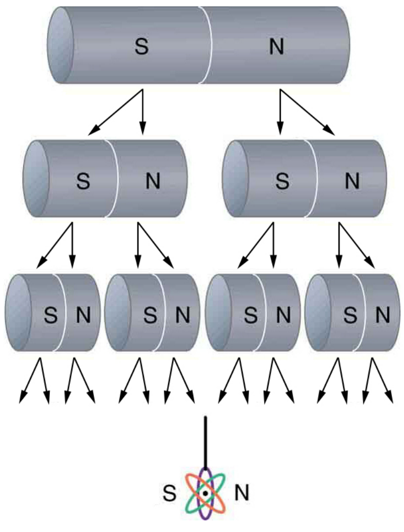
Misconception Alert: Earth's Geographic North Pole Hides an S

The Earth acts like a very large bar magnet with its south-seeking pole near the geographic North Pole. That is why the north pole of your compass is attracted toward the geographic north pole of the Earth—because the magnetic pole that is near the geographic North Pole is actually a south magnetic pole! Confusion arises because the geographic term "North Pole" has come to be used (incorrectly) for the magnetic pole that is near the

North Pole. Thus, “North magnetic pole” is actually a misnomer—it should be called the South magnetic pole.



Unlike poles attract, whereas like poles repel.



North and south poles always occur in pairs. Attempts

to separate them result in more pairs of poles. If we continue to split the magnet, we will eventually get down to an iron atom with a north pole and a south pole—these, too, cannot be separated.

The fact that magnetic poles always occur in pairs of north and south is true from the very large scale—for example, sunspots always occur in pairs that are north and south magnetic poles—all the way down to the very small scale. Magnetic atoms have both a north pole and a south pole, as do many types of subatomic particles, such as electrons, protons, and neutrons.

Note:
Making Connections: Take-Home Experiment—Refrigerator Magnets
We know that like magnetic poles repel and unlike poles attract. See if you can show this for two refrigerator magnets. Will the magnets stick if you turn them over? Why do they stick to the door anyway? What can you say about the magnetic properties of the door next to the magnet? Do refrigerator magnets stick to metal or plastic spoons? Do they stick to all types of metal?

Section Summary

- Magnetism is a subject that includes the properties of magnets, the effect of the magnetic force on moving charges and currents, and the

creation of magnetic fields by currents.
- There are two types of magnetic poles, called the north magnetic pole and south magnetic pole.
- North magnetic poles are those that are attracted toward the Earth's geographic north pole.
- Like poles repel and unlike poles attract.
- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.

Conceptual Questions

Exercise:

Problem:

Volcanic and other such activity at the mid-Atlantic ridge extrudes material to fill the gap between separating tectonic plates associated with continental drift. The magnetization of rocks is found to reverse in a coordinated manner with distance from the ridge. What does this imply about the Earth's magnetic field and how could the knowledge of the spreading rate be used to give its historical record?

Glossary

north magnetic pole
: the end or the side of a magnet that is attracted toward Earth's geographic north pole

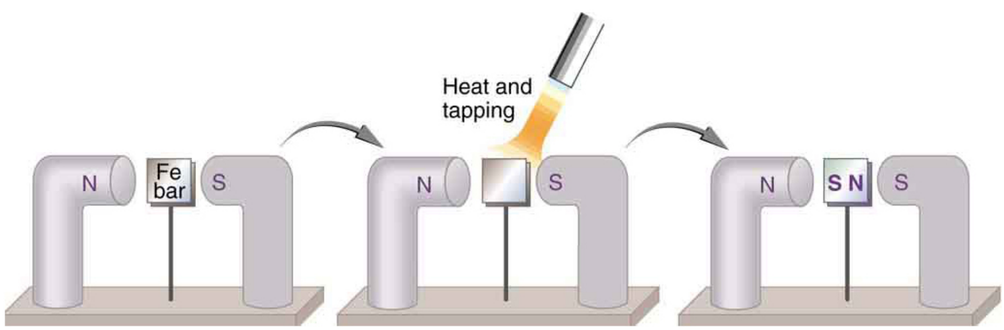
south magnetic pole
: the end or the side of a magnet that is attracted toward Earth's geographic south pole

Ferromagnets and Electromagnets

- Define ferromagnet.
- Describe the role of magnetic domains in magnetization.
- Explain the significance of the Curie temperature.
- Describe the relationship between electricity and magnetism.

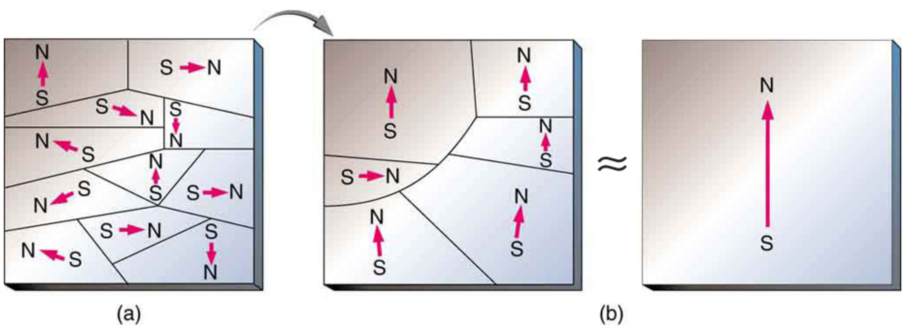
Ferromagnets

Only certain materials, such as iron, cobalt, nickel, and gadolinium, exhibit strong magnetic effects. Such materials are called **ferromagnetic**, after the Latin word for iron, *ferrum*. A group of materials made from the alloys of the rare earth elements are also used as strong and permanent magnets; a popular one is neodymium. Other materials exhibit weak magnetic effects, which are detectable only with sensitive instruments. Not only do ferromagnetic materials respond strongly to magnets (the way iron is attracted to magnets), they can also be **magnetized** themselves—that is, they can be induced to be magnetic or made into permanent magnets.



An unmagnetized piece of iron is placed between two magnets, heated, and then cooled, or simply tapped when cold. The iron becomes a permanent magnet with the poles aligned as shown: its south pole is adjacent to the north pole of the original magnet, and its north pole is adjacent to the south pole of the original magnet. Note that there are attractive forces between the magnets.

When a magnet is brought near a previously unmagnetized ferromagnetic material, it causes local magnetization of the material with unlike poles closest, as in [link]. (This results in the attraction of the previously unmagnetized material to the magnet.) What happens on a microscopic scale is illustrated in [link]. The regions within the material called **domains** act like small bar magnets. Within domains, the poles of individual atoms are aligned. Each atom acts like a tiny bar magnet. Domains are small and randomly oriented in an unmagnetized ferromagnetic object. In response to an external magnetic field, the domains may grow to millimeter size, aligning themselves as shown in [link](b). This induced magnetization can be made permanent if the material is heated and then cooled, or simply tapped in the presence of other magnets.



(a) An unmagnetized piece of iron (or other ferromagnetic material) has randomly oriented domains. (b) When magnetized by an external field, the domains show greater alignment, and some grow at the expense of others. Individual atoms are aligned within domains; each atom acts like a tiny bar magnet.

Conversely, a permanent magnet can be demagnetized by hard blows or by heating it in the absence of another magnet. Increased thermal motion at higher temperature can disrupt and randomize the orientation and the size of

the domains. There is a well-defined temperature for ferromagnetic materials, which is called the **Curie temperature**, above which they cannot be magnetized. The Curie temperature for iron is 1043 K $(770^{\circ}C)$, which is well above room temperature. There are several elements and alloys that have Curie temperatures much lower than room temperature and are ferromagnetic only below those temperatures.

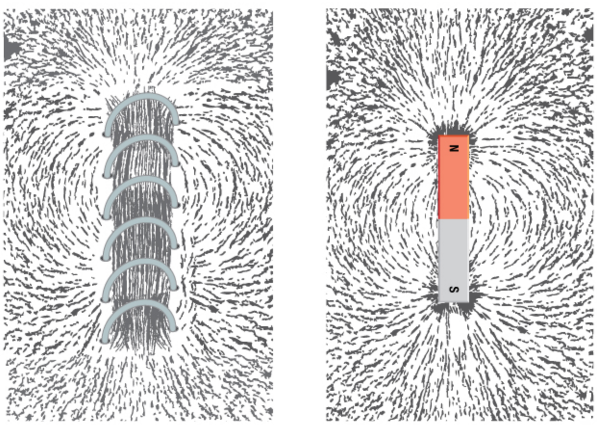
Electromagnets

Early in the 19th century, it was discovered that electrical currents cause magnetic effects. The first significant observation was by the Danish scientist Hans Christian Oersted (1777–1851), who found that a compass needle was deflected by a current-carrying wire. This was the first significant evidence that the movement of charges had any connection with magnets. **Electromagnetism** is the use of electric current to make magnets. These temporarily induced magnets are called **electromagnets**. Electromagnets are employed for everything from a wrecking yard crane that lifts scrapped cars to controlling the beam of a 90-km-circumference particle accelerator to the magnets in medical imaging machines (See [link]).

Instrument for magnetic resonance imaging (MRI). The device uses a superconducting

cylindrical coil for the main magnetic field. The patient goes into this “tunnel” on the gurney. (credit: Bill McChesney, Flickr)

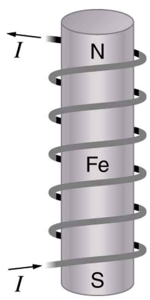
[link] shows that the response of iron filings to a current-carrying coil and to a permanent bar magnet. The patterns are similar. In fact, electromagnets and ferromagnets have the same basic characteristics—for example, they have north and south poles that cannot be separated and for which like poles repel and unlike poles attract.



(a) (b)

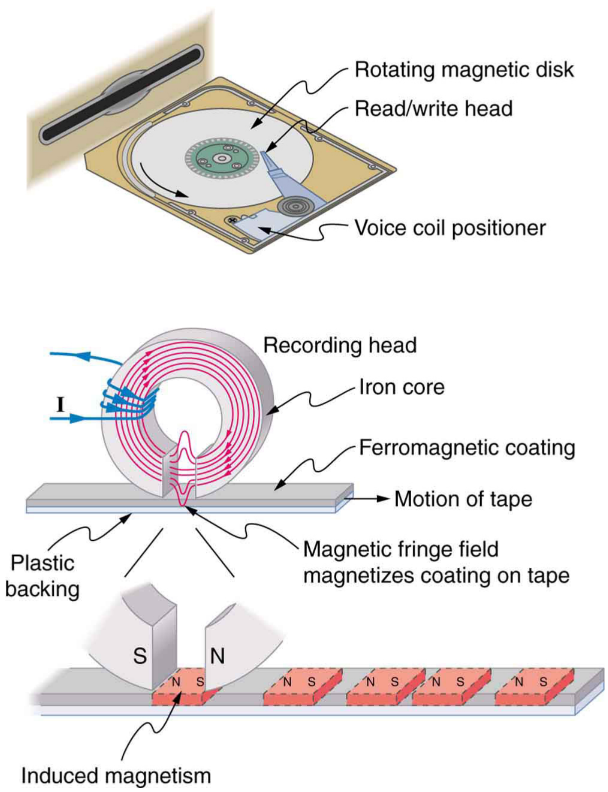
Iron filings near (a) a current-carrying coil and (b) a magnet act like tiny compass needles, showing the shape of their fields. Their response to a current-carrying coil and a permanent magnet is seen to be very similar, especially near the ends of the coil and the magnet.

Combining a ferromagnet with an electromagnet can produce particularly strong magnetic effects. (See [link].) Whenever strong magnetic effects are needed, such as lifting scrap metal, or in particle accelerators, electromagnets are enhanced by ferromagnetic materials. Limits to how strong the magnets can be made are imposed by coil resistance (it will overheat and melt at sufficiently high current), and so superconducting magnets may be employed. These are still limited, because superconducting properties are destroyed by too great a magnetic field.



An electromagnet with a ferromagnetic core can produce very strong magnetic effects. Alignment of domains in the core produces a magnet, the poles of which are aligned with the electromagnet.

[link] shows a few uses of combinations of electromagnets and ferromagnets. Ferromagnetic materials can act as memory devices, because the orientation of the magnetic fields of small domains can be reversed or erased. Magnetic information storage on videotapes and computer hard drives are among the most common applications. This property is vital in our digital world.



An electromagnet induces regions of permanent magnetism on a floppy disk coated with a ferromagnetic material. The information stored here is digital (a region is either magnetic or not); in other applications, it can be analog

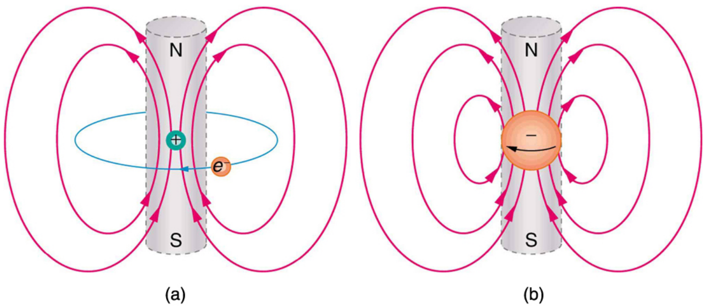
(with a varying strength), such
as on audiotapes.

Current: The Source of All Magnetism

An electromagnet creates magnetism with an electric current. In later sections we explore this more quantitatively, finding the strength and direction of magnetic fields created by various currents. But what about ferromagnets? [link] shows models of how electric currents create magnetism at the submicroscopic level. (Note that we cannot directly observe the paths of individual electrons about atoms, and so a model or visual image, consistent with all direct observations, is made. We can directly observe the electron's orbital angular momentum, its spin momentum, and subsequent magnetic moments, all of which are explained with electric-current-creating subatomic magnetism.) Currents, including those associated with other submicroscopic particles like protons, allow us to explain ferromagnetism and all other magnetic effects. Ferromagnetism, for example, results from an internal cooperative alignment of electron spins, possible in some materials but not in others.

Crucial to the statement that electric current is the source of all magnetism is the fact that it is impossible to separate north and south magnetic poles. (This is far different from the case of positive and negative charges, which are easily separated.) A current loop always produces a magnetic dipole—that is, a magnetic field that acts like a north pole and south pole pair. Since isolated north and south magnetic poles, called **magnetic monopoles**, are not observed, currents are used to explain all magnetic effects. If magnetic monopoles did exist, then we would have to modify this underlying connection that all magnetism is due to electrical current. There is no known reason that magnetic monopoles should not exist—they are simply never observed—and so searches at the subnuclear level continue. If they do *not* exist, we would like to find out why not. If they *do* exist, we would like to see evidence of them.

Note:
Electric Currents and Magnetism
Electric current is the source of all magnetism.



(a) In the planetary model of the atom, an electron orbits a nucleus, forming a closed-current loop and producing a magnetic field with a north pole and a south pole. (b) Electrons have spin and can be crudely pictured as rotating charge, forming a current that produces a magnetic field with a north pole and a south pole. Neither the planetary model nor the image of a spinning electron is completely consistent with modern physics. However, they do provide a useful way of understanding phenomena.

Note:
PhET Explorations: Magnets and Electromagnets
Explore the interactions between a compass and bar magnet. Discover how you can use a battery and wire to make a magnet! Can you make it a stronger magnet? Can you make the magnetic field reverse?

https://archive.cnx.org/specials/92176000-ae74-11e5-baad-cfab91c15075/magnets-and-electromagnets/#sim-bar-magnet

Section Summary

- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.
- All magnetism is created by electric current.
- Ferromagnetic materials, such as iron, are those that exhibit strong magnetic effects.
- The atoms in ferromagnetic materials act like small magnets (due to currents within the atoms) and can be aligned, usually in millimeter-sized regions called domains.
- Domains can grow and align on a larger scale, producing permanent magnets. Such a material is magnetized, or induced to be magnetic.
- Above a material's Curie temperature, thermal agitation destroys the alignment of atoms, and ferromagnetism disappears.
- Electromagnets employ electric currents to make magnetic fields, often aided by induced fields in ferromagnetic materials.

Glossary

ferromagnetic
: materials, such as iron, cobalt, nickel, and gadolinium, that exhibit strong magnetic effects

magnetized
: to be turned into a magnet; to be induced to be magnetic

domains
: regions within a material that behave like small bar magnets

Curie temperature
: the temperature above which a ferromagnetic material cannot be magnetized

electromagnetism
: the use of electrical currents to induce magnetism

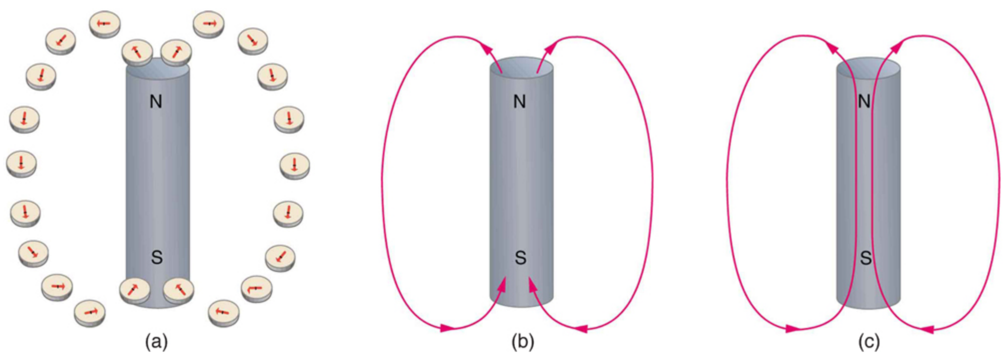
electromagnet
: an object that is temporarily magnetic when an electrical current is passed through it

magnetic monopoles
: an isolated magnetic pole; a south pole without a north pole, or vice versa (no magnetic monopole has ever been observed)

Magnetic Fields and Magnetic Field Lines

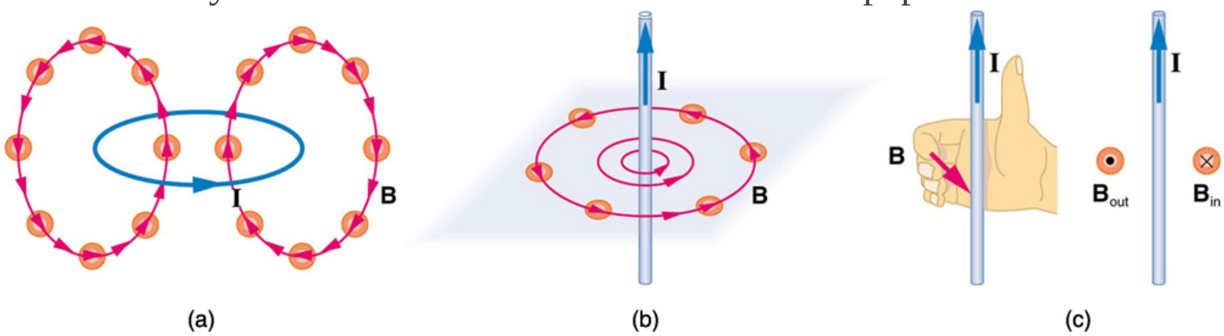
- Define magnetic field and describe the magnetic field lines of various magnetic fields.

Einstein is said to have been fascinated by a compass as a child, perhaps musing on how the needle felt a force without direct physical contact. His ability to think deeply and clearly about action at a distance, particularly for gravitational, electric, and magnetic forces, later enabled him to create his revolutionary theory of relativity. Since magnetic forces act at a distance, we define a **magnetic field** to represent magnetic forces. The pictorial representation of **magnetic field lines** is very useful in visualizing the strength and direction of the magnetic field. As shown in [link], the **direction of magnetic field lines** is defined to be the direction in which the north end of a compass needle points. The magnetic field is traditionally called the ***B*-field**.



Magnetic field lines are defined to have the direction that a small compass points when placed at a location. (a) If small compasses are used to map the magnetic field around a bar magnet, they will point in the directions shown: away from the north pole of the magnet, toward the south pole of the magnet. (Recall that the Earth's north magnetic pole is really a south pole in terms of definitions of poles on a bar magnet.) (b) Connecting the arrows gives continuous magnetic field lines. The strength of the field is proportional to the closeness (or density) of the lines. (c) If the interior of the magnet could be probed, the field lines would be found to form continuous closed loops.

Small compasses used to test a magnetic field will not disturb it. (This is analogous to the way we tested electric fields with a small test charge. In both cases, the fields represent only the object creating them and not the probe testing them.) [link] shows how the magnetic field appears for a current loop and a long straight wire, as could be explored with small compasses. A small compass placed in these fields will align itself parallel to the field line at its location, with its north pole pointing in the direction of *B*. Note the symbols used for field into and out of the paper.



Small compasses could be used to map the fields shown here. (a) The magnetic field of a circular current loop is similar to that of a bar magnet. (b) A long and straight wire creates a field with magnetic field lines forming circular loops. (c) When the wire is in the plane of the paper, the field is perpendicular to the paper. Note that the symbols used for the field pointing inward (like the tail of an arrow) and the field pointing outward (like the tip of an arrow).

Note:
Making Connections: Concept of a Field
A field is a way of mapping forces surrounding any object that can act on another object at a distance without apparent physical connection. The field represents the object generating it. Gravitational fields map

gravitational forces, electric fields map electrical forces, and magnetic fields map magnetic forces.

Extensive exploration of magnetic fields has revealed a number of hard-and-fast rules. We use magnetic field lines to represent the field (the lines are a pictorial tool, not a physical entity in and of themselves). The properties of magnetic field lines can be summarized by these rules:

1. The direction of the magnetic field is tangent to the field line at any point in space. A small compass will point in the direction of the field line.
2. The strength of the field is proportional to the closeness of the lines. It is exactly proportional to the number of lines per unit area perpendicular to the lines (called the areal density).
3. Magnetic field lines can never cross, meaning that the field is unique at any point in space.
4. Magnetic field lines are continuous, forming closed loops without beginning or end. They go from the north pole to the south pole.

The last property is related to the fact that the north and south poles cannot be separated. It is a distinct difference from electric field lines, which begin and end on the positive and negative charges. If magnetic monopoles existed, then magnetic field lines would begin and end on them.

Section Summary

- Magnetic fields can be pictorially represented by magnetic field lines, the properties of which are as follows:

1. The field is tangent to the magnetic field line.
2. Field strength is proportional to the line density.
3. Field lines cannot cross.
4. Field lines are continuous loops.

Conceptual Questions

Exercise:

Problem:

Explain why the magnetic field would not be unique (that is, not have a single value) at a point in space where magnetic field lines might cross. (Consider the direction of the field at such a point.)

Exercise:

Problem:

List the ways in which magnetic field lines and electric field lines are similar. For example, the field direction is tangent to the line at any point in space. Also list the ways in which they differ. For example, electric force is parallel to electric field lines, whereas magnetic force on moving charges is perpendicular to magnetic field lines.

Exercise:

Problem:

Noting that the magnetic field lines of a bar magnet resemble the electric field lines of a pair of equal and opposite charges, do you expect the magnetic field to rapidly decrease in strength with distance from the magnet? Is this consistent with your experience with magnets?

Exercise:

Problem:

Is the Earth's magnetic field parallel to the ground at all locations? If not, where is it parallel to the surface? Is its strength the same at all locations? If not, where is it greatest?

Glossary

magnetic field
: the representation of magnetic forces

B-field
another term for magnetic field

magnetic field lines
the pictorial representation of the strength and the direction of a magnetic field

direction of magnetic field lines
the direction that the north end of a compass needle points

Magnetic Field Strength: Force on a Moving Charge in a Magnetic Field

- Describe the effects of magnetic fields on moving charges.
- Use the right hand rule 1 to determine the velocity of a charge, the direction of the magnetic field, and the direction of the magnetic force on a moving charge.
- Calculate the magnetic force on a moving charge.

What is the mechanism by which one magnet exerts a force on another? The answer is related to the fact that all magnetism is caused by current, the flow of charge. *Magnetic fields exert forces on moving charges*, and so they exert forces on other magnets, all of which have moving charges.

Right Hand Rule 1

The magnetic force on a moving charge is one of the most fundamental known. Magnetic force is as important as the electrostatic or Coulomb force. Yet the magnetic force is more complex, in both the number of factors that affects it and in its direction, than the relatively simple Coulomb force. The magnitude of the **magnetic force** F on a charge q moving at a speed v in a magnetic field of strength B is given by
Equation:

$$F = \text{qvB}\ \text{sin}\ \theta,$$

where θ is the angle between the directions of $\mathbf{v}$ and $\mathbf{B}$. This force is often called the **Lorentz force**. In fact, this is how we define the magnetic field strength B—in terms of the force on a charged particle moving in a magnetic field. The SI unit for magnetic field strength B is called the **tesla** (T) after the eccentric but brilliant inventor Nikola Tesla (1856–1943). To determine how the tesla relates to other SI units, we solve $F = \text{qvB}\ \text{sin}\ \theta$ for B.
Equation:

$$B = \frac{F}{\text{qv}\ \text{sin}\ \theta}$$

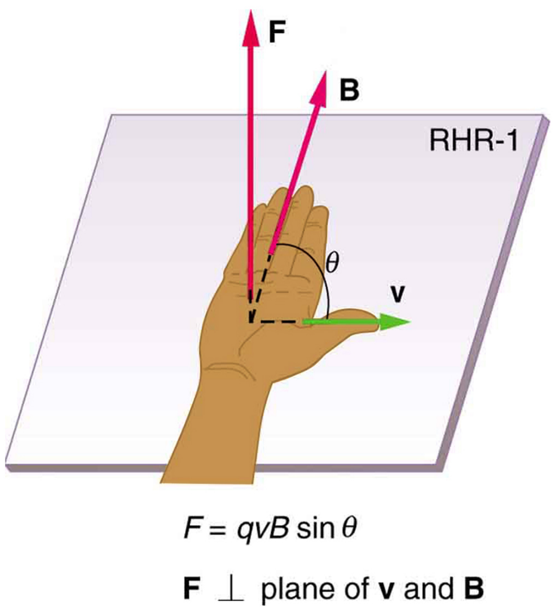
Because sin θ is unitless, the tesla is
Equation:

$$1 \text{ T} = \frac{1 \text{ N}}{\text{C} \cdot \text{m/s}} = \frac{1 \text{ N}}{\text{A} \cdot \text{m}}$$

(note that C/s = A).

Another smaller unit, called the **gauss** (G), where $1 \text{ G} = 10^{-4} \text{ T}$, is sometimes used. The strongest permanent magnets have fields near 2 T; superconducting electromagnets may attain 10 T or more. The Earth's magnetic field on its surface is only about $5 \times 10^{-5} \text{ T}$, or 0.5 G.

The *direction* of the magnetic force $\mathbf{F}$ is perpendicular to the plane formed by $\mathbf{v}$ and $\mathbf{B}$, as determined by the **right hand rule 1** (or RHR-1), which is illustrated in [link]. RHR-1 states that, to determine the direction of the magnetic force on a positive moving charge, you point the thumb of the right hand in the direction of $\mathbf{v}$, the fingers in the direction of $\mathbf{B}$, and a perpendicular to the palm points in the direction of $\mathbf{F}$. One way to remember this is that there is one velocity, and so the thumb represents it. There are many field lines, and so the fingers represent them. The force is in the direction you would push with your palm. The force on a negative charge is in exactly the opposite direction to that on a positive charge.



Magnetic fields exert forces on moving charges. This force is one of the most basic known. The direction of the magnetic force on a moving charge is perpendicular to the plane formed by $\mathbf{v}$ and $\mathbf{B}$ and follows right hand rule–1 (RHR-1) as shown. The magnitude of the force is proportional to q, v, B, and the sine of the angle between $\mathbf{v}$ and $\mathbf{B}$.

Note:

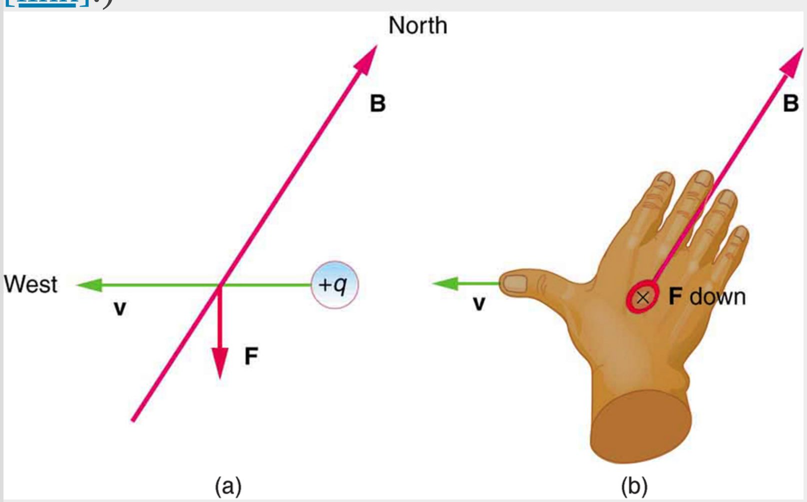
Making Connections: Charges and Magnets

There is no magnetic force on static charges. However, there is a magnetic force on moving charges. When charges are stationary, their electric fields do not affect magnets. But, when charges move, they produce magnetic

fields that exert forces on other magnets. When there is relative motion, a connection between electric and magnetic fields emerges—each affects the other.

Example:
Calculating Magnetic Force: Earth's Magnetic Field on a Charged Glass Rod

With the exception of compasses, you seldom see or personally experience forces due to the Earth's small magnetic field. To illustrate this, suppose that in a physics lab you rub a glass rod with silk, placing a 20-nC positive charge on it. Calculate the force on the rod due to the Earth's magnetic field, if you throw it with a horizontal velocity of 10 m/s due west in a place where the Earth's field is due north parallel to the ground. (The direction of the force is determined with right hand rule 1 as shown in [link].)



A positively charged object moving due west in a region where the Earth's magnetic field is due north experiences a force that is straight down as shown. A negative charge moving in the same direction would feel a force straight up.

Strategy

We are given the charge, its velocity, and the magnetic field strength and direction. We can thus use the equation $F = \text{qvB sin } \theta$ to find the force.

Solution

The magnetic force is

Equation:

$$F = \text{qvb sin } \theta.$$

We see that $\text{sin } \theta = 1$, since the angle between the velocity and the direction of the field is 90º. Entering the other given quantities yields

Equation:

$$\begin{array}{lll} F & = & \left(20 \times 10^{-9} \text{ C}\right)\left(10 \text{ m/s}\right)\left(5 \times 10^{-5} \text{ T}\right) \\ & = & 1 \times 10^{-11} \left(\text{C} \cdot \text{m/s}\right)\left(\frac{\text{N}}{\text{C} \cdot \text{m/s}}\right) = 1 \times 10^{-11} \text{ N.} \end{array}$$

Discussion

This force is completely negligible on any macroscopic object, consistent with experience. (It is calculated to only one digit, since the Earth's field varies with location and is given to only one digit.) The Earth's magnetic field, however, does produce very important effects, particularly on submicroscopic particles. Some of these are explored in Force on a Moving Charge in a Magnetic Field: Examples and Applications.

Section Summary

- Magnetic fields exert a force on a moving charge q, the magnitude of which is

  Equation:

  $$F = \text{qvB sin } \theta,$$

  where θ is the angle between the directions of v and B.
- The SI unit for magnetic field strength B is the tesla (T), which is related to other units by

Equation:

$$1\text{ T}=\frac{1\text{ N}}{\text{C}\cdot\text{m/s}}=\frac{1\text{ N}}{\text{A}\cdot\text{m}}.$$

- The *direction* of the force on a moving charge is given by right hand rule 1 (RHR-1): Point the thumb of the right hand in the direction of v, the fingers in the direction of B, and a perpendicular to the palm points in the direction of F.
- The force is perpendicular to the plane formed by $\mathbf{v}$ and $\mathbf{B}$. Since the force is zero if $\mathbf{v}$ is parallel to $\mathbf{B}$, charged particles often follow magnetic field lines rather than cross them.

Conceptual Questions

Exercise:

Problem:

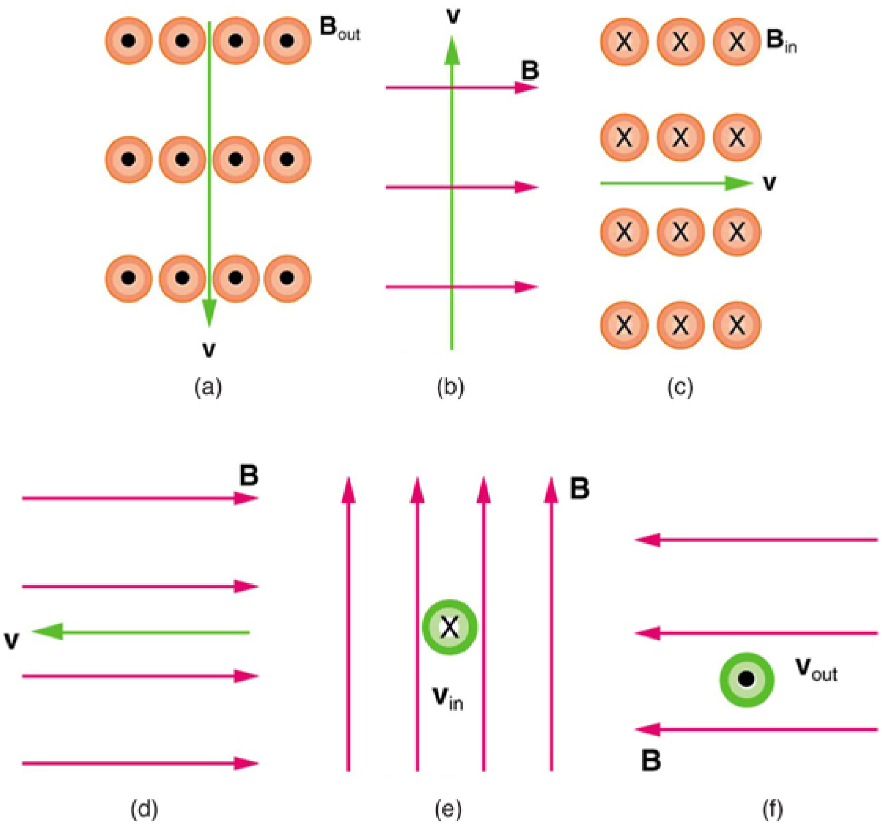
If a charged particle moves in a straight line through some region of space, can you say that the magnetic field in that region is necessarily zero?

Problems & Exercises

Exercise:

Problem:

What is the direction of the magnetic force on a positive charge that moves as shown in each of the six cases shown in [link]?



Solution:

(a) Left (West)

(b) Into the page

(c) Up (North)

(d) No force

(e) Right (East)

(f) Down (South)

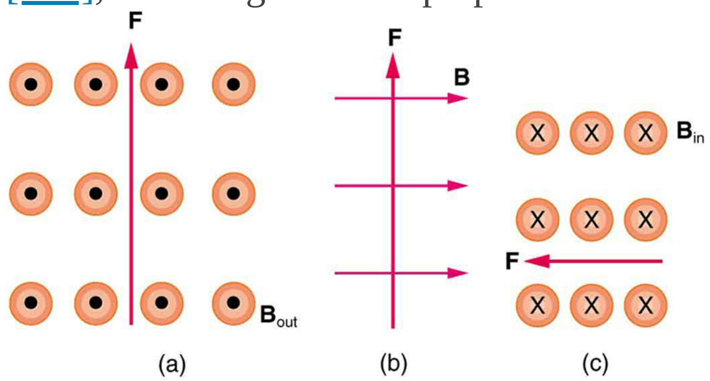
Exercise:

Problem: Repeat [link] for a negative charge.

Exercise:

Problem:

What is the direction of the velocity of a negative charge that experiences the magnetic force shown in each of the three cases in [link], assuming it moves perpendicular to $\mathbf{B}$?



Solution:

(a) East (right)

(b) Into page

(c) South (down)

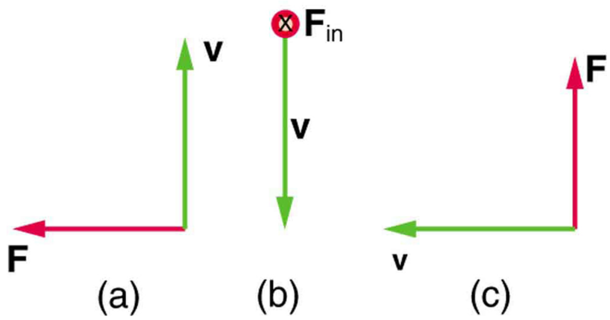
Exercise:

Problem: Repeat [link] for a positive charge.

Exercise:

Problem:

What is the direction of the magnetic field that produces the magnetic force on a positive charge as shown in each of the three cases in the figure below, assuming $\mathbf{B}$ is perpendicular to $\mathbf{v}$?



Solution:

(a) Into page

(b) West (left)

(c) Out of page

Exercise:

Problem: Repeat [link] for a negative charge.

Exercise:

Problem:

What is the maximum magnitude of the force on an aluminum rod with a $0.100\text{-}\mu\text{C}$ charge that you pass between the poles of a 1.50-T permanent magnet at a speed of 5.00 m/s? In what direction is the force?

Solution:

7.50×10^{-7} N perpendicular to both the magnetic field lines and the velocity

Exercise:

Problem:

(a) Aircraft sometimes acquire small static charges. Suppose a supersonic jet has a $0.500\text{-}\mu\text{C}$ charge and flies due west at a speed of 660 m/s over the Earth’s magnetic south pole (near Earth's geographic north pole), where the $8.00 \times 10^{-5}\text{-T}$ magnetic field points straight down. What are the direction and the magnitude of the magnetic force on the plane? (b) Discuss whether the value obtained in part (a) implies this is a significant or negligible effect.

Exercise:

Problem:

(a) A cosmic ray proton moving toward the Earth at $5.00 \times 10^{7}\ \text{m/s}$ experiences a magnetic force of $1.70 \times 10^{-16}\ \text{N}$. What is the strength of the magnetic field if there is a 45° angle between it and the proton’s velocity? (b) Is the value obtained in part (a) consistent with the known strength of the Earth’s magnetic field on its surface? Discuss.

Solution:

(a) $3.01 \times 10^{-5}\ \text{T}$

(b) This is slightly less then the magnetic field strength of $5 \times 10^{-5}\ \text{T}$ at the surface of the Earth, so it is consistent.

Exercise:

Problem:

An electron moving at $4.00 \times 10^{3}\ \text{m/s}$ in a 1.25-T magnetic field experiences a magnetic force of $1.40 \times 10^{-16}\ \text{N}$. What angle does the velocity of the electron make with the magnetic field? There are two answers.

Exercise:

Problem:

(a) A physicist performing a sensitive measurement wants to limit the magnetic force on a moving charge in her equipment to less than 1.00×10^{-12} N. What is the greatest the charge can be if it moves at a maximum speed of 30.0 m/s in the Earth's field? (b) Discuss whether it would be difficult to limit the charge to less than the value found in (a) by comparing it with typical static electricity and noting that static is often absent.

Solution:

(a) 6.67×10^{-10} C (taking the Earth's field to be 5.00×10^{-5} T)

(b) Less than typical static, therefore difficult

Glossary

right hand rule 1 (RHR-1)
: the rule to determine the direction of the magnetic force on a positive moving charge: when the thumb of the right hand points in the direction of the charge's velocity **v** and the fingers point in the direction of the magnetic field $\mathbf{B}$, then the force on the charge is perpendicular and away from the palm; the force on a negative charge is perpendicular and into the palm

Lorentz force
: the force on a charge moving in a magnetic field

tesla
: T, the SI unit of the magnetic field strength; $1\ \mathrm{T} = \frac{1\ \mathrm{N}}{\mathrm{A \cdot m}}$

magnetic force
: the force on a charge produced by its motion through a magnetic field; the Lorentz force

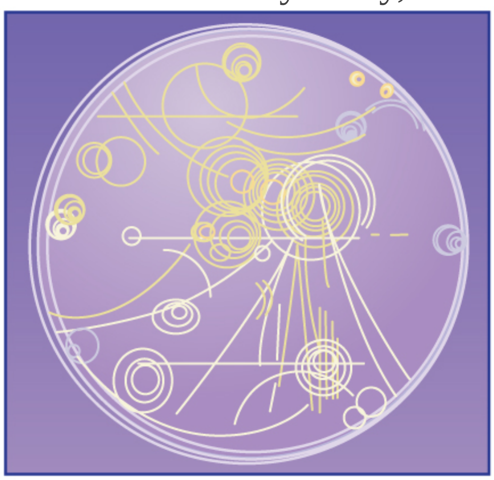
gauss

G, the unit of the magnetic field strength; $1\ \mathrm{G} = 10^{-4}\ \mathrm{T}$

Force on a Moving Charge in a Magnetic Field: Examples and Applications

- Describe the effects of a magnetic field on a moving charge.
- Calculate the radius of curvature of the path of a charge that is moving in a magnetic field.

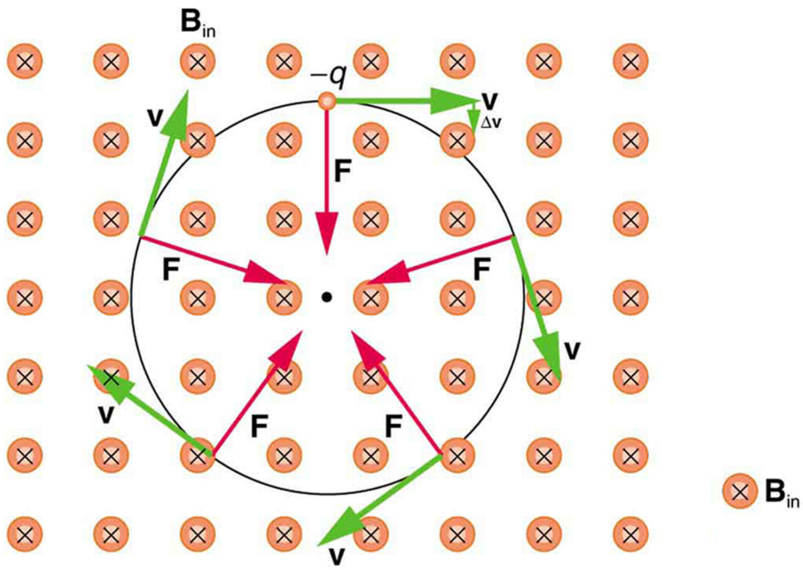
Magnetic force can cause a charged particle to move in a circular or spiral path. Cosmic rays are energetic charged particles in outer space, some of which approach the Earth. They can be forced into spiral paths by the Earth's magnetic field. Protons in giant accelerators are kept in a circular path by magnetic force. The bubble chamber photograph in [link] shows charged particles moving in such curved paths. The curved paths of charged particles in magnetic fields are the basis of a number of phenomena and can even be used analytically, such as in a mass spectrometer.



Trails of bubbles are produced by high-energy charged particles moving through the superheated liquid hydrogen in this artist's rendition of a bubble chamber. There is a strong magnetic field perpendicular to the page that causes the curved paths of the particles. The radius of the path can be

used to find the mass, charge, and energy of the particle.

So does the magnetic force cause circular motion? Magnetic force is always perpendicular to velocity, so that it does no work on the charged particle. The particle's kinetic energy and speed thus remain constant. The direction of motion is affected, but not the speed. This is typical of uniform circular motion. The simplest case occurs when a charged particle moves perpendicular to a uniform B-field, such as shown in [link]. (If this takes place in a vacuum, the magnetic field is the dominant factor determining the motion.) Here, the magnetic force supplies the centripetal force $F_{\mathrm{c}} = mv^2/r$. Noting that $\sin \theta = 1$, we see that $F = \mathrm{qvB}$.



A negatively charged particle moves in the plane of the page in a region where the magnetic field is perpendicular into the page (represented by the small circles with x's—like the tails of arrows). The magnetic force is perpendicular to the velocity, and so velocity changes in direction but not magnitude. Uniform circular motion results.

Because the magnetic force F supplies the centripetal force F_c, we have
Equation:

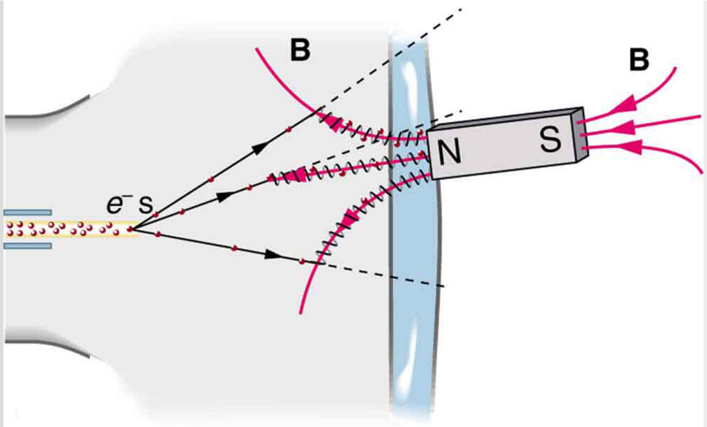
$$\text{qvB} = \frac{mv^2}{r}.$$

Solving for r yields
Equation:

$$r = \frac{\text{mv}}{\text{qB}}.$$

Here, r is the radius of curvature of the path of a charged particle with mass m and charge q, moving at a speed v perpendicular to a magnetic field of strength B. If the velocity is not perpendicular to the magnetic field, then v is the component of the velocity perpendicular to the field. The component of the velocity parallel to the field is unaffected, since the magnetic force is zero for motion parallel to the field. This produces a spiral motion rather than a circular one.

Example:
Calculating the Curvature of the Path of an Electron Moving in a Magnetic Field: A Magnet on a TV Screen
A magnet brought near an old-fashioned TV screen such as in [link] (TV sets with cathode ray tubes instead of LCD screens) severely distorts its picture by altering the path of the electrons that make its phosphors glow. ***(Don't try this at home, as it will permanently magnetize and ruin the TV.)*** To illustrate this, calculate the radius of curvature of the path of an electron having a velocity of 6.00×10^7 m/s (corresponding to the accelerating voltage of about 10.0 kV used in some TVs) perpendicular to a magnetic field of strength $B = 0.500$ T (obtainable with permanent magnets).



Side view showing what happens when a magnet comes in contact with a computer monitor or TV screen. Electrons moving toward the screen spiral about magnetic field lines, maintaining the component of their velocity parallel to the field lines. This distorts the image on the screen.

Strategy

We can find the radius of curvature r directly from the equation $r = \frac{mv}{qB}$, since all other quantities in it are given or known.

Solution

Using known values for the mass and charge of an electron, along with the given values of v and B gives us

Equation:

$$\begin{aligned} r = \frac{\text{mv}}{\text{qB}} &= \frac{\left(9.11\times10^{-31}\text{ kg}\right)\left(6.00\times10^{7}\text{ m/s}\right)}{\left(1.60\times10^{-19}\text{ C}\right)(0.500\text{ T})} \\ &= 6.83\times10^{-4}\text{ m} \end{aligned}$$

or

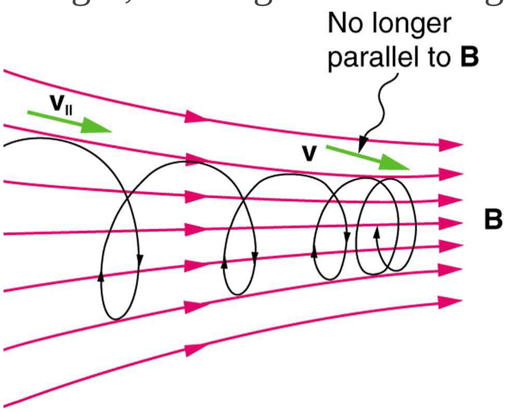
Equation:

$$r = 0.683\text{ mm.}$$

Discussion

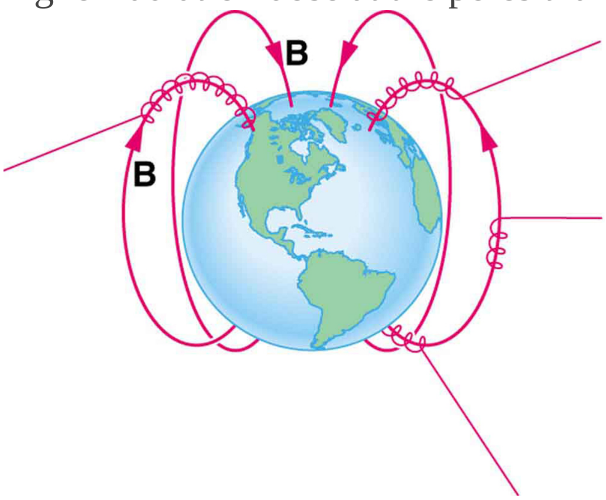
The small radius indicates a large effect. The electrons in the TV picture tube are made to move in very tight circles, greatly altering their paths and distorting the image.

[link] shows how electrons not moving perpendicular to magnetic field lines follow the field lines. The component of velocity parallel to the lines is unaffected, and so the charges spiral along the field lines. If field strength increases in the direction of motion, the field will exert a force to slow the charges, forming a kind of magnetic mirror, as shown below.



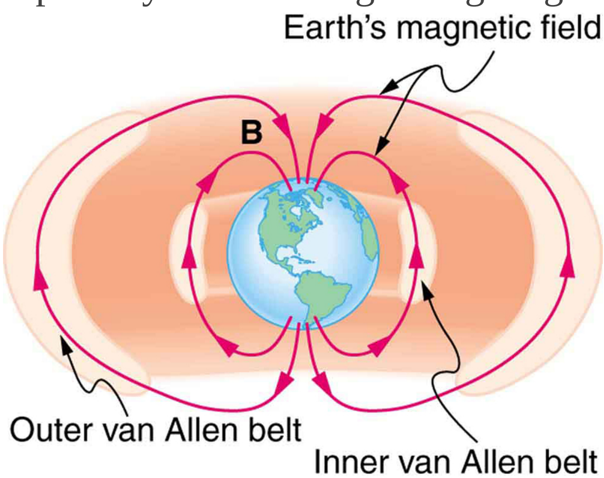
When a charged particle moves along a magnetic field line into a region where the field becomes stronger, the particle experiences a force that reduces the component of velocity parallel to the field. This force slows the motion along the field line and here reverses it, forming a “magnetic mirror.”

The properties of charged particles in magnetic fields are related to such different things as the Aurora Australis or Aurora Borealis and particle accelerators. *Charged particles approaching magnetic field lines may get trapped in spiral orbits about the lines rather than crossing them*, as seen above. Some cosmic rays, for example, follow the Earth's magnetic field lines, entering the atmosphere near the magnetic poles and causing the southern or northern lights through their ionization of molecules in the atmosphere. This glow of energized atoms and molecules is seen in [link]. Those particles that approach middle latitudes must cross magnetic field lines, and many are prevented from penetrating the atmosphere. Cosmic rays are a component of background radiation; consequently, they give a higher radiation dose at the poles than at the equator.



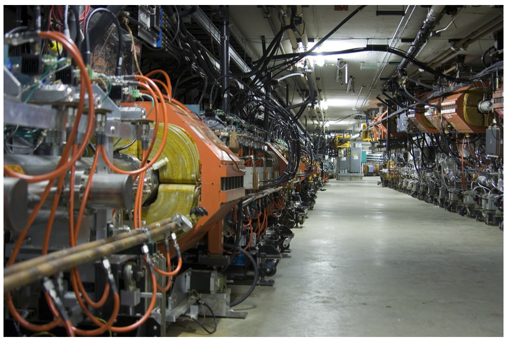
Energetic electrons and protons, components of cosmic rays, from the Sun and deep outer space often follow the Earth's magnetic field lines rather than cross them. (Recall that the Earth's north magnetic pole is really a south pole in terms of a bar magnet.)

Some incoming charged particles become trapped in the Earth’s magnetic field, forming two belts above the atmosphere known as the Van Allen radiation belts after the discoverer James A. Van Allen, an American astrophysicist. (See [link].) Particles trapped in these belts form radiation fields (similar to nuclear radiation) so intense that manned space flights avoid them and satellites with sensitive electronics are kept out of them. In the few minutes it took lunar missions to cross the Van Allen radiation belts, astronauts received radiation doses more than twice the allowed annual exposure for radiation workers. Other planets have similar belts, especially those having strong magnetic fields like Jupiter.



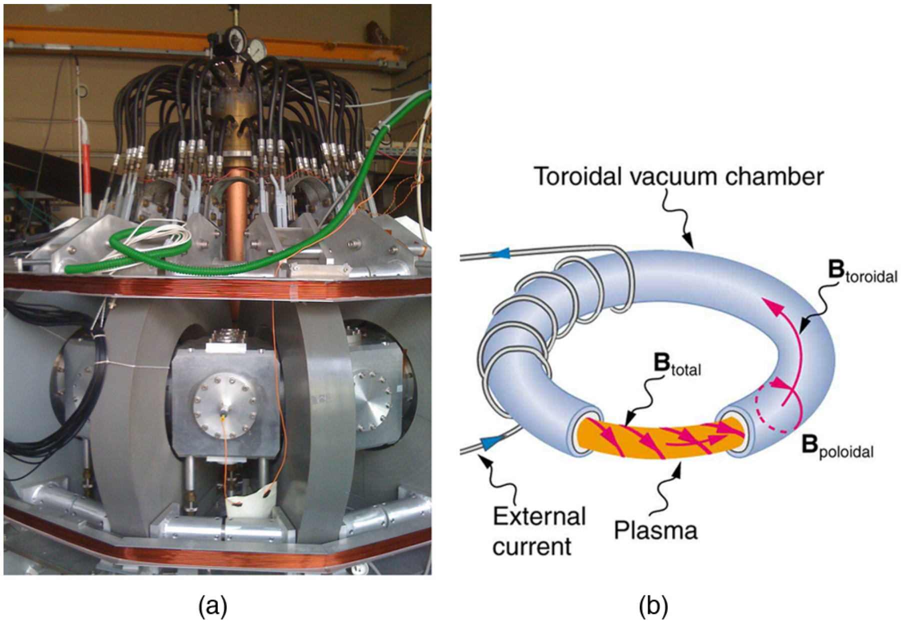
The Van Allen radiation belts are two regions in which energetic charged particles are trapped in the Earth’s magnetic field. One belt lies about 300 km above the Earth’s surface, the other about 16,000 km. Charged particles in these belts migrate along magnetic field lines and are partially reflected away from the poles by the stronger fields there. The charged particles that enter the atmosphere are replenished by the Sun and sources in deep outer space.

Back on Earth, we have devices that employ magnetic fields to contain charged particles. Among them are the giant particle accelerators that have been used to explore the substructure of matter. (See [link].) Magnetic fields not only control the direction of the charged particles, they also are used to focus particles into beams and overcome the repulsion of like charges in these beams.



The Fermilab facility in Illinois has a large particle accelerator (the most powerful in the world until 2008) that employs magnetic fields (magnets seen here in orange) to contain and direct its beam. This and other accelerators have been in use for several decades and have allowed us to discover some of the laws underlying all matter. (credit: ammcrim, Flickr)

Thermonuclear fusion (like that occurring in the Sun) is a hope for a future clean energy source. One of the most promising devices is the *tokamak*, which uses magnetic fields to contain (or trap) and direct the reactive charged particles. (See [link].) Less exotic, but more immediately practical, amplifiers in microwave ovens use a magnetic field to contain oscillating electrons. These oscillating electrons generate the microwaves sent into the oven.



Tokamaks such as the one shown in the figure are being studied with the goal of economical production of energy by nuclear fusion. Magnetic fields in the doughnut-shaped device contain and direct the reactive charged particles. (credit: David Mellis, Flickr)

Mass spectrometers have a variety of designs, and many use magnetic fields to measure mass. The curvature of a charged particle's path in the field is related to its mass and is measured to obtain mass information. (See More Applications of Magnetism.) Historically, such techniques were employed in the first direct observations of electron charge and mass. Today, mass

spectrometers (sometimes coupled with gas chromatographs) are used to determine the make-up and sequencing of large biological molecules.

Section Summary

- Magnetic force can supply centripetal force and cause a charged particle to move in a circular path of radius
  Equation:
  $$r = \frac{\text{mv}}{\text{qB}}$$
  where v is the component of the velocity perpendicular to B for a charged particle with mass m and charge q.

Conceptual Questions

Exercise:

Problem:

How can the motion of a charged particle be used to distinguish between a magnetic and an electric field?

Exercise:

Problem:

High-velocity charged particles can damage biological cells and are a component of radiation exposure in a variety of locations ranging from research facilities to natural background. Describe how you could use a magnetic field to shield yourself.

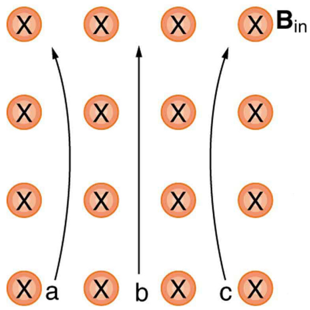
Exercise:

Problem:

If a cosmic ray proton approaches the Earth from outer space along a line toward the center of the Earth that lies in the plane of the equator, in what direction will it be deflected by the Earth's magnetic field? What about an electron? A neutron?

Exercise:

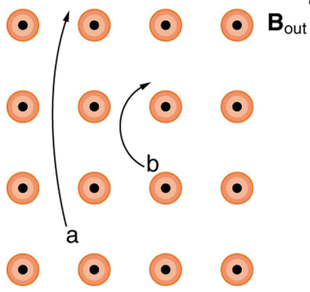
Problem: What are the signs of the charges on the particles in [link]?



Exercise:

Problem:

Which of the particles in [link] has the greatest velocity, assuming they have identical charges and masses?



Exercise:

Problem:

Which of the particles in [link] has the greatest mass, assuming all have identical charges and velocities?

Exercise:

Problem:

While operating, a high-precision TV monitor is placed on its side during maintenance. The image on the monitor changes color and blurs slightly. Discuss the possible relation of these effects to the Earth's magnetic field.

Problems & Exercises

If you need additional support for these problems, see More Applications of Magnetism.

Exercise:

Problem:

A cosmic ray electron moves at $7.50 \times 10^{6}\ \mathrm{m/s}$ perpendicular to the Earth's magnetic field at an altitude where field strength is $1.00 \times 10^{-5}\ \mathrm{T}$. What is the radius of the circular path the electron follows?

Solution:

4.27 m

Exercise:

Problem:

A proton moves at $7.50 \times 10^{7}\ \mathrm{m/s}$ perpendicular to a magnetic field. The field causes the proton to travel in a circular path of radius 0.800 m. What is the field strength?

Exercise:

Problem:

(a) Viewers of *Star Trek* hear of an antimatter drive on the Starship *Enterprise*. One possibility for such a futuristic energy source is to store antimatter charged particles in a vacuum chamber, circulating in a magnetic field, and then extract them as needed. Antimatter annihilates with normal matter, producing pure energy. What strength magnetic field is needed to hold antiprotons, moving at $5.00 \times 10^7\ \text{m/s}$ in a circular path 2.00 m in radius? Antiprotons have the same mass as protons but the opposite (negative) charge. (b) Is this field strength obtainable with today's technology or is it a futuristic possibility?

Solution:

(a) 0.261 T

(b) This strength is definitely obtainable with today's technology. Magnetic field strengths of 0.500 T are obtainable with permanent magnets.

Exercise:

Problem:

(a) An oxygen-16 ion with a mass of $2.66 \times 10^{-26}\ \text{kg}$ travels at $5.00 \times 10^6\ \text{m/s}$ perpendicular to a 1.20-T magnetic field, which makes it move in a circular arc with a 0.231-m radius. What positive charge is on the ion? (b) What is the ratio of this charge to the charge of an electron? (c) Discuss why the ratio found in (b) should be an integer.

Exercise:

Problem:

What radius circular path does an electron travel if it moves at the same speed and in the same magnetic field as the proton in [link]?

Solution:

$4.36 \times 10^{-4}\ \mathrm{m}$

Exercise:

Problem:

A velocity selector in a mass spectrometer uses a 0.100-T magnetic field. (a) What electric field strength is needed to select a speed of $4.00 \times 10^{6}\ \mathrm{m/s}$? (b) What is the voltage between the plates if they are separated by 1.00 cm?

Exercise:

Problem:

An electron in a TV CRT moves with a speed of $6.00 \times 10^{7}\ \mathrm{m/s}$, in a direction perpendicular to the Earth's field, which has a strength of $5.00 \times 10^{-5}\ \mathrm{T}$. (a) What strength electric field must be applied perpendicular to the Earth's field to make the electron moves in a straight line? (b) If this is done between plates separated by 1.00 cm, what is the voltage applied? (Note that TVs are usually surrounded by a ferromagnetic material to shield against external magnetic fields and avoid the need for such a correction.)

Solution:

(a) 3.00 kV/m

(b) 30.0 V

Exercise:

Problem:

(a) At what speed will a proton move in a circular path of the same radius as the electron in [link]? (b) What would the radius of the path be if the proton had the same speed as the electron? (c) What would the radius be if the proton had the same kinetic energy as the electron? (d) The same momentum?

Exercise:

Problem:

A mass spectrometer is being used to separate common oxygen-16 from the much rarer oxygen-18, taken from a sample of old glacial ice. (The relative abundance of these oxygen isotopes is related to climatic temperature at the time the ice was deposited.) The ratio of the masses of these two ions is 16 to 18, the mass of oxygen-16 is 2.66×10^{-26} kg, and they are singly charged and travel at 5.00×10^{6} m/s in a 1.20-T magnetic field. What is the separation between their paths when they hit a target after traversing a semicircle?

Solution:

0.173 m

Exercise:

Problem:

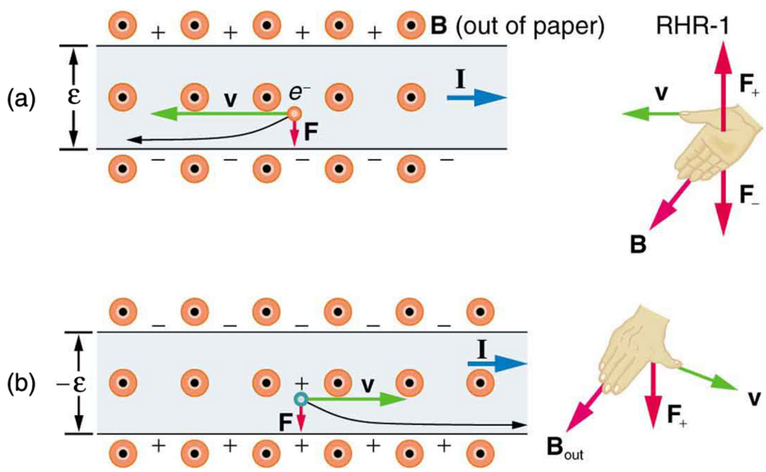
(a) Triply charged uranium-235 and uranium-238 ions are being separated in a mass spectrometer. (The much rarer uranium-235 is used as reactor fuel.) The masses of the ions are 3.90×10^{-25} kg and 3.95×10^{-25} kg, respectively, and they travel at 3.00×10^{5} m/s in a 0.250-T field. What is the separation between their paths when they hit a target after traversing a semicircle? (b) Discuss whether this distance between their paths seems to be big enough to be practical in the separation of uranium-235 from uranium-238.

The Hall Effect

- Describe the Hall effect.
- Calculate the Hall emf across a current-carrying conductor.

We have seen effects of a magnetic field on free-moving charges. The magnetic field also affects charges moving in a conductor. One result is the Hall effect, which has important implications and applications.

[link] shows what happens to charges moving through a conductor in a magnetic field. The field is perpendicular to the electron drift velocity and to the width of the conductor. Note that conventional current is to the right in both parts of the figure. In part (a), electrons carry the current and move to the left. In part (b), positive charges carry the current and move to the right. Moving electrons feel a magnetic force toward one side of the conductor, leaving a net positive charge on the other side. This separation of charge *creates a voltage* ε, known as the **Hall emf**, *across* the conductor. The creation of a voltage *across* a current-carrying conductor by a magnetic field is known as the **Hall effect**, after Edwin Hall, the American physicist who discovered it in 1879.



The Hall effect. (a) Electrons move to the left in this flat conductor (conventional current to the right). The magnetic field is directly out of the page, represented by circled dots; it exerts a force on the moving charges, causing a voltage ε, the

Hall emf, across the conductor. (b) Positive charges moving to the right (conventional current also to the right) are moved to the side, producing a Hall emf of the opposite sign, –ε. Thus, if the direction of the field and current are known, the sign of the charge carriers can be determined from the Hall effect.

One very important use of the Hall effect is to determine whether positive or negative charges carries the current. Note that in [link](b), where positive charges carry the current, the Hall emf has the sign opposite to when negative charges carry the current. Historically, the Hall effect was used to show that electrons carry current in metals and it also shows that positive charges carry current in some semiconductors. The Hall effect is used today as a research tool to probe the movement of charges, their drift velocities and densities, and so on, in materials. In 1980, it was discovered that the Hall effect is quantized, an example of quantum behavior in a macroscopic object.

The Hall effect has other uses that range from the determination of blood flow rate to precision measurement of magnetic field strength. To examine these quantitatively, we need an expression for the Hall emf, ε, across a conductor. Consider the balance of forces on a moving charge in a situation where B, v, and l are mutually perpendicular, such as shown in [link]. Although the magnetic force moves negative charges to one side, they cannot build up without limit. The electric field caused by their separation opposes the magnetic force, $F = \text{qvB}$, and the electric force, $F_{\text{e}} = \text{qE}$, eventually grows to equal it. That is,

Equation:

$$\text{qE} = \text{qvB}$$

or

Equation:

$$E = \text{vB}.$$

Note that the electric field E is uniform across the conductor because the magnetic field B is uniform, as is the conductor. For a uniform electric field, the relationship between electric field and voltage is $E = \varepsilon / l$, where l is the width of the conductor and ε is the Hall emf. Entering this into the last expression gives

Equation:

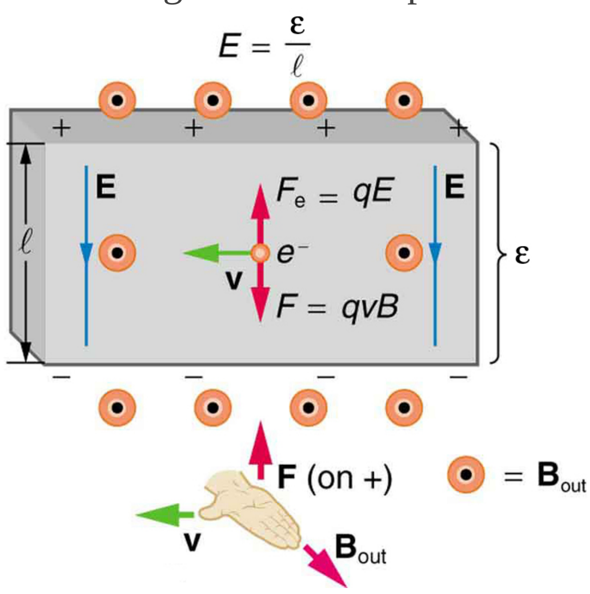
$$\frac{\varepsilon}{l} = \text{vB}.$$

Solving this for the Hall emf yields

Equation:

$$\varepsilon = Blv \; (B \;\; v \;\text{ and } l \;\text{ mutually perpendicular})$$

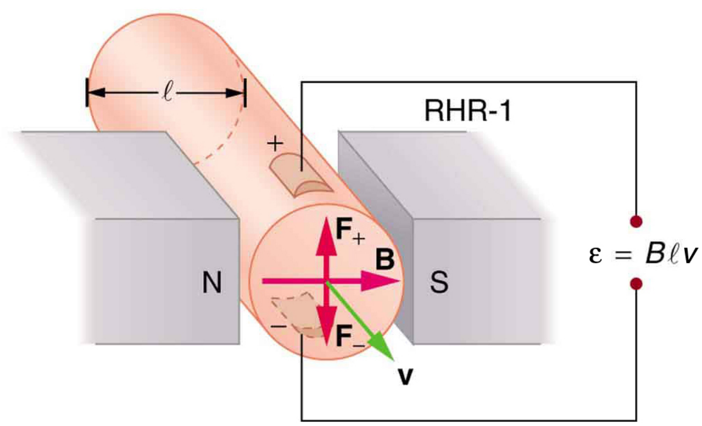
where ε is the Hall effect voltage across a conductor of width l through which charges move at a speed v.



The Hall emf ε produces an electric force that balances the magnetic force on the moving

charges. The magnetic force produces charge separation, which builds up until it is balanced by the electric force, an equilibrium that is quickly reached.

One of the most common uses of the Hall effect is in the measurement of magnetic field strength B. Such devices, called *Hall probes*, can be made very small, allowing fine position mapping. Hall probes can also be made very accurate, usually accomplished by careful calibration. Another application of the Hall effect is to measure fluid flow in any fluid that has free charges (most do). (See [link].) A magnetic field applied perpendicular to the flow direction produces a Hall emf ε as shown. Note that the sign of ε depends not on the sign of the charges, but only on the directions of B and v. The magnitude of the Hall emf is $\varepsilon = \text{Blv}$, where l is the pipe diameter, so that the average velocity v can be determined from ε providing the other factors are known.



The Hall effect can be used to measure fluid flow in any fluid having free charges, such as blood. The Hall emf ε is measured across the tube perpendicular to the applied magnetic field and is proportional to the average velocity v.

Example:
Calculating the Hall emf: Hall Effect for Blood Flow
A Hall effect flow probe is placed on an artery, applying a 0.100-T magnetic field across it, in a setup similar to that in [link]. What is the Hall emf, given the vessel's inside diameter is 4.00 mm and the average blood velocity is 20.0 cm/s?

Strategy
Because B, v, and l are mutually perpendicular, the equation $\varepsilon = \text{Blv}$ can be used to find ε.

Solution
Entering the given values for B, v, and l gives

Equation:

$$\begin{array}{lll}\varepsilon & = & Blv = (0.100\text{ T})\left(4.00\times 10^{-3}\text{ m}\right)(0.200\text{ m/s})\\ & = & 80.0\text{ μV}\end{array}$$

Discussion
This is the average voltage output. Instantaneous voltage varies with pulsating blood flow. The voltage is small in this type of measurement. ε is particularly difficult to measure, because there are voltages associated with heart action (ECG voltages) that are on the order of millivolts. In practice, this difficulty is overcome by applying an AC magnetic field, so that the Hall emf is AC with the same frequency. An amplifier can be very selective in picking out only the appropriate frequency, eliminating signals and noise at other frequencies.

Section Summary

- The Hall effect is the creation of voltage ε, known as the Hall emf, across a current-carrying conductor by a magnetic field.
- The Hall emf is given by

Equation:

$$\varepsilon = Blv\,(B\ \ v\ \ \text{and}\ l\ \ \text{mutually perpendicular})$$

for a conductor of width l through which charges move at a speed v.

Conceptual Questions

Exercise:

Problem:

Discuss how the Hall effect could be used to obtain information on free charge density in a conductor. (Hint: Consider how drift velocity and current are related.)

Problems & Exercises

Exercise:

Problem:

A large water main is 2.50 m in diameter and the average water velocity is 6.00 m/s. Find the Hall voltage produced if the pipe runs perpendicular to the Earth's $5.00 \times 10^{-5}\text{-T}$ field.

Solution:

$7.50 \times 10^{-4}\ \text{V}$

Exercise:

Problem:

What Hall voltage is produced by a 0.200-T field applied across a 2.60-cm-diameter aorta when blood velocity is 60.0 cm/s?

Exercise:

Problem:

(a) What is the speed of a supersonic aircraft with a 17.0-m wingspan, if it experiences a 1.60-V Hall voltage between its wing tips when in level flight over the north magnetic pole, where the Earth’s field strength is 8.00×10^{-5} T? (b) Explain why very little current flows as a result of this Hall voltage.

Solution:

(a) 1.18×10^{3} m/s

(b) Once established, the Hall emf pushes charges one direction and the magnetic force acts in the opposite direction resulting in no net force on the charges. Therefore, no current flows in the direction of the Hall emf. This is the same as in a current-carrying conductor—current does not flow in the direction of the Hall emf.

Exercise:

Problem:

A nonmechanical water meter could utilize the Hall effect by applying a magnetic field across a metal pipe and measuring the Hall voltage produced. What is the average fluid velocity in a 3.00-cm-diameter pipe, if a 0.500-T field across it creates a 60.0-mV Hall voltage?

Exercise:

Problem:

Calculate the Hall voltage induced on a patient’s heart while being scanned by an MRI unit. Approximate the conducting path on the heart wall by a wire 7.50 cm long that moves at 10.0 cm/s perpendicular to a 1.50-T magnetic field.

Solution:

11.3 mV

Exercise:

Problem:

A Hall probe calibrated to read 1.00 μV when placed in a 2.00-T field is placed in a 0.150-T field. What is its output voltage?

Exercise:

Problem:

Using information in [link], what would the Hall voltage be if a 2.00-T field is applied across a 10-gauge copper wire (2.588 mm in diameter) carrying a 20.0-A current?

Solution:

1.16 μV

Exercise:

Problem:

Show that the Hall voltage across wires made of the same material, carrying identical currents, and subjected to the same magnetic field is inversely proportional to their diameters. (Hint: Consider how drift velocity depends on wire diameter.)

Exercise:

Problem:

A patient with a pacemaker is mistakenly being scanned for an MRI image. A 10.0-cm-long section of pacemaker wire moves at a speed of 10.0 cm/s perpendicular to the MRI unit's magnetic field and a 20.0-mV Hall voltage is induced. What is the magnetic field strength?

Solution:

2.00 T

Glossary

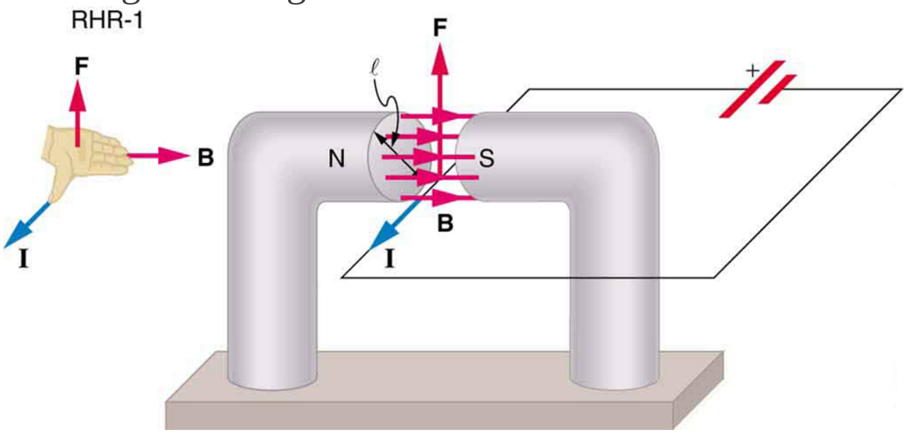
Hall effect
: the creation of voltage across a current-carrying conductor by a magnetic field

Hall emf
: the electromotive force created by a current-carrying conductor by a magnetic field, $\varepsilon = \mathrm{B}lv$

Magnetic Force on a Current-Carrying Conductor

- Describe the effects of a magnetic force on a current-carrying conductor.
- Calculate the magnetic force on a current-carrying conductor.

Because charges ordinarily cannot escape a conductor, the magnetic force on charges moving in a conductor is transmitted to the conductor itself.



The magnetic field exerts a force on a current-carrying wire in a direction given by the right hand rule 1 (the same direction as that on the individual moving charges). This force can easily be large enough to move the wire, since typical currents consist of very large numbers of moving charges.

We can derive an expression for the magnetic force on a current by taking a sum of the magnetic forces on individual charges. (The forces add because they are in the same direction.) The force on an individual charge moving at the drift velocity v_{d} is given by $F = qv_{\mathrm{d}}B \sin \theta$. Taking B to be uniform over a length of wire l and zero elsewhere, the total magnetic force on the wire is then $F = (qv_{\mathrm{d}}B \sin \theta)(N)$, where N is the number of charge carriers in the section of wire of length l. Now, $N = \mathrm{nV}$, where n is the number of charge carriers per unit volume and V is the volume of wire in the field. Noting that $V = \mathrm{Al}$, where A is the cross-sectional area of the

wire, then the force on the wire is $F = (qv_\text{d}B\sin\theta)(\text{nAl})$. Gathering terms,

Equation:

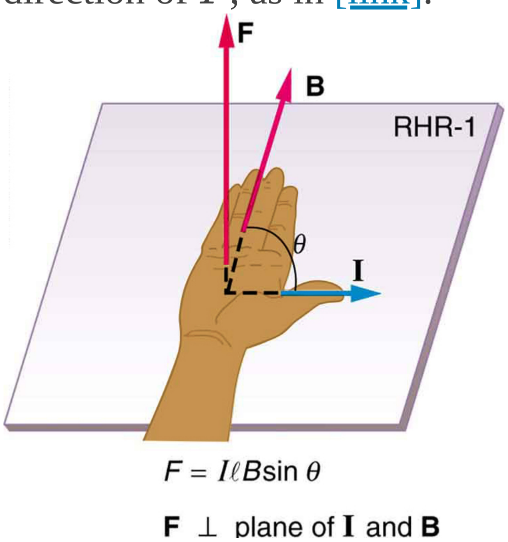
$$F = (nqAv_\text{d})lB\sin\theta.$$

Because $nqAv_\text{d} = I$ (see Current),

Equation:

$$F = \text{IlB}\sin\theta$$

is the equation for *magnetic force on a length* l *of wire carrying a current* I *in a uniform magnetic field* B, as shown in [link]. If we divide both sides of this expression by l, we find that the magnetic force per unit length of wire in a uniform field is $\frac{F}{l} = \text{IB}\sin\theta$. The direction of this force is given by RHR-1, with the thumb in the direction of the current I. Then, with the fingers in the direction of B, a perpendicular to the palm points in the direction of F, as in [link].



The force on a current-carrying wire in a magnetic field is $F = \text{IlB}\sin\theta$. Its

direction is given by
RHR-1.

Example:
Calculating Magnetic Force on a Current-Carrying Wire: A Strong Magnetic Field
Calculate the force on the wire shown in [link], given $B = 1.50\text{ T}$, $l = 5.00\text{ cm}$, and $I = 20.0\text{ A}$.
Strategy
The force can be found with the given information by using $F = \text{IlB sin }\theta$ and noting that the angle θ between I and B is 90º, so that $\text{sin }\theta = 1$.
Solution
Entering the given values into $F = \text{IlB sin }\theta$ yields
Equation:

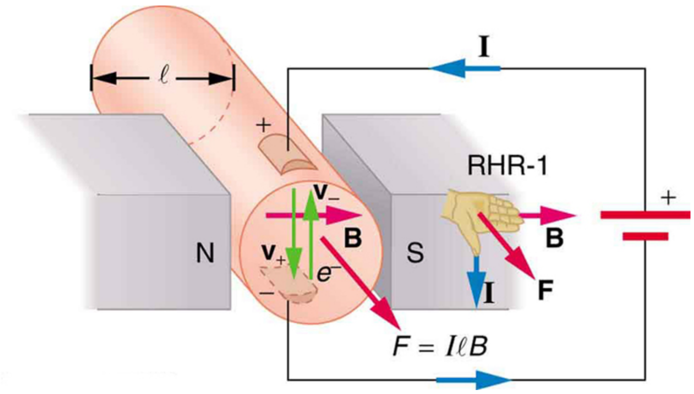
$$F = \text{IlB sin }\theta = (20.0\text{ A})(0.0500\text{ m})(1.50\text{ T})(1).$$

The units for tesla are $1\text{ T} = \frac{\text{N}}{\text{A}\cdot\text{m}}$; thus,
Equation:

$$F = 1.50\text{ N}.$$

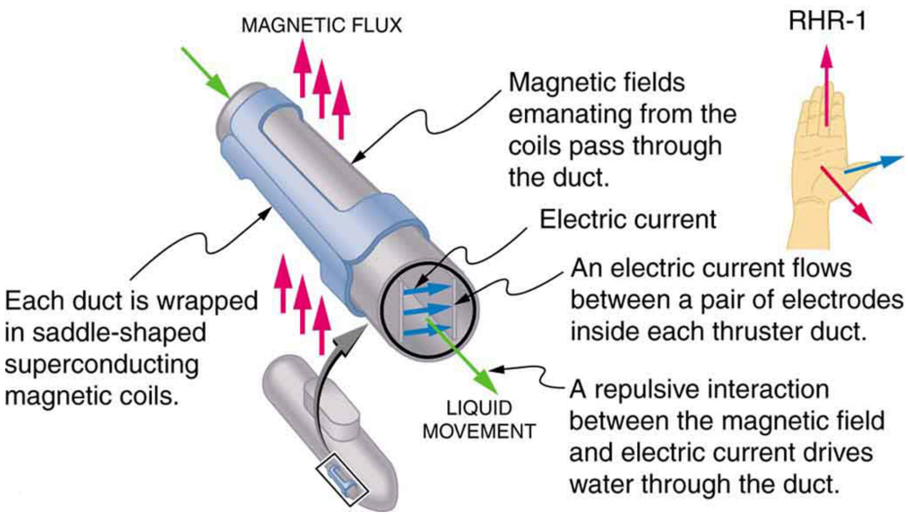
Discussion
This large magnetic field creates a significant force on a small length of wire.

Magnetic force on current-carrying conductors is used to convert electric energy to work. (Motors are a prime example—they employ loops of wire and are considered in the next section.) Magnetohydrodynamics (MHD) is the technical name given to a clever application where magnetic force pumps fluids without moving mechanical parts. (See [link].)



Magnetohydrodynamics. The magnetic force on the current passed through this fluid can be used as a nonmechanical pump.

A strong magnetic field is applied across a tube and a current is passed through the fluid at right angles to the field, resulting in a force on the fluid parallel to the tube axis as shown. The absence of moving parts makes this attractive for moving a hot, chemically active substance, such as the liquid sodium employed in some nuclear reactors. Experimental artificial hearts are testing with this technique for pumping blood, perhaps circumventing the adverse effects of mechanical pumps. (Cell membranes, however, are affected by the large fields needed in MHD, delaying its practical application in humans.) MHD propulsion for nuclear submarines has been proposed, because it could be considerably quieter than conventional propeller drives. The deterrent value of nuclear submarines is based on their ability to hide and survive a first or second nuclear strike. As we slowly disassemble our nuclear weapons arsenals, the submarine branch will be the last to be decommissioned because of this ability (See [link].) Existing MHD drives are heavy and inefficient—much development work is needed.



An MHD propulsion system in a nuclear submarine could produce significantly less turbulence than propellers and allow it to run more silently. The development of a silent drive submarine was dramatized in the book and the film *The Hunt for Red October*.

Section Summary

- The magnetic force on current-carrying conductors is given by **Equation:**

$$F = \text{IlB sin } \theta,$$

where I is the current, l is the length of a straight conductor in a uniform magnetic field B, and θ is the angle between I and B. The force follows RHR-1 with the thumb in the direction of I.

Conceptual Questions

Exercise:

Problem:

Draw a sketch of the situation in [link] showing the direction of electrons carrying the current, and use RHR-1 to verify the direction of the force on the wire.

Exercise:

Problem:

Verify that the direction of the force in an MHD drive, such as that in [link], does not depend on the sign of the charges carrying the current across the fluid.

Exercise:

Problem:

Why would a magnetohydrodynamic drive work better in ocean water than in fresh water? Also, why would superconducting magnets be desirable?

Exercise:

Problem:

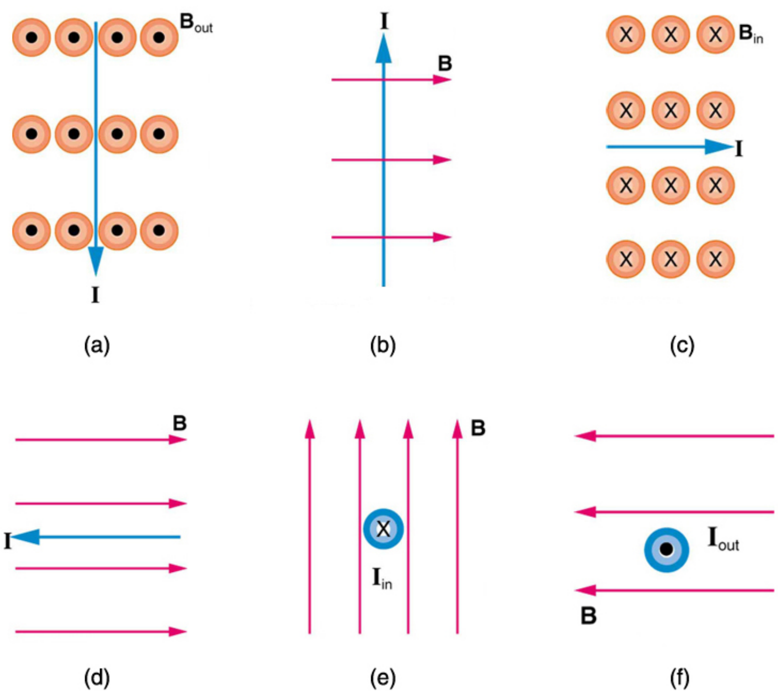
Which is more likely to interfere with compass readings, AC current in your refrigerator or DC current when you start your car? Explain.

Problems & Exercises

Exercise:

Problem:

What is the direction of the magnetic force on the current in each of the six cases in [link]?



Solution:

(a) west (left)

(b) into page

(c) north (up)

(d) no force

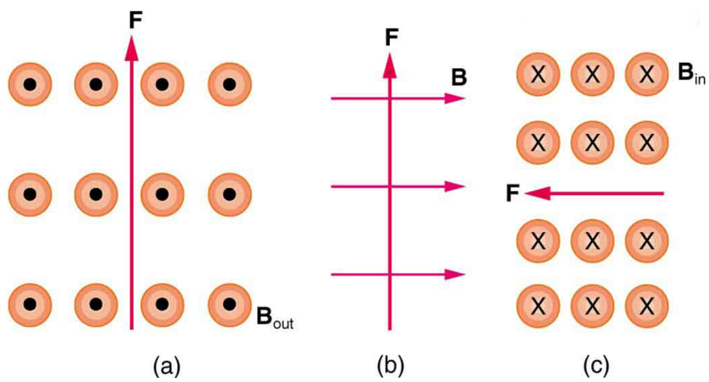
(e) east (right)

(f) south (down)

Exercise:

Problem:

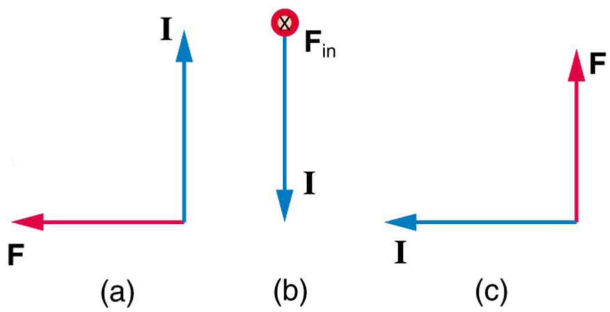
What is the direction of a current that experiences the magnetic force shown in each of the three cases in [link], assuming the current runs perpendicular to B?



Exercise:

Problem:

What is the direction of the magnetic field that produces the magnetic force shown on the currents in each of the three cases in [link], assuming is perpendicular to ?



Solution:

(a) into page

(b) west (left)

(c) out of page

Exercise:

Problem:

(a) What is the force per meter on a lightning bolt at the equator that carries 20,000 A perpendicular to the Earth’s $3.00 \times 10^{-5}\text{-T}$ field? (b) What is the direction of the force if the current is straight up and the Earth’s field direction is due north, parallel to the ground?

Exercise:

Problem:

(a) A DC power line for a light-rail system carries 1000 A at an angle of 30.0^{o} to the Earth’s 5.00×10^{-5} -T field. What is the force on a 100-m section of this line? (b) Discuss practical concerns this presents, if any.

Solution:

(a) 2.50 N

(b) This is about half a pound of force per 100 m of wire, which is much less than the weight of the wire itself. Therefore, it does not cause any special concerns.

Exercise:

Problem:

What force is exerted on the water in an MHD drive utilizing a 25.0-cm-diameter tube, if 100-A current is passed across the tube that is perpendicular to a 2.00-T magnetic field? (The relatively small size of this force indicates the need for very large currents and magnetic fields to make practical MHD drives.)

Exercise:

Problem:

A wire carrying a 30.0-A current passes between the poles of a strong magnet that is perpendicular to its field and experiences a 2.16-N force on the 4.00 cm of wire in the field. What is the average field strength?

Solution:

1.80 T

Exercise:

Problem:

(a) A 0.750-m-long section of cable carrying current to a car starter motor makes an angle of 60° with the Earth's 5.50×10^{-5} T field. What is the current when the wire experiences a force of 7.00×10^{-3} N? (b) If you run the wire between the poles of a strong horseshoe magnet, subjecting 5.00 cm of it to a 1.75-T field, what force is exerted on this segment of wire?

Exercise:

Problem:

(a) What is the angle between a wire carrying an 8.00-A current and the 1.20-T field it is in if 50.0 cm of the wire experiences a magnetic force of 2.40 N? (b) What is the force on the wire if it is rotated to make an angle of 90° with the field?

Solution:

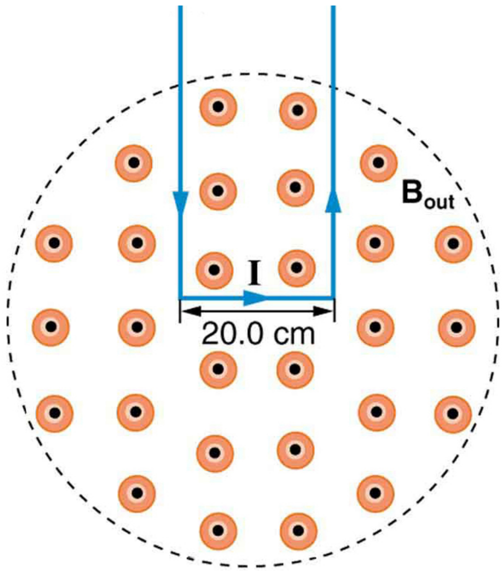
(a) 30°

(b) 4.80 N

Exercise:

Problem:

The force on the rectangular loop of wire in the magnetic field in [link] can be used to measure field strength. The field is uniform, and the plane of the loop is perpendicular to the field. (a) What is the direction of the magnetic force on the loop? Justify the claim that the forces on the sides of the loop are equal and opposite, independent of how much of the loop is in the field and do not affect the net force on the loop. (b) If a current of 5.00 A is used, what is the force per tesla on the 20.0-cm-wide loop?

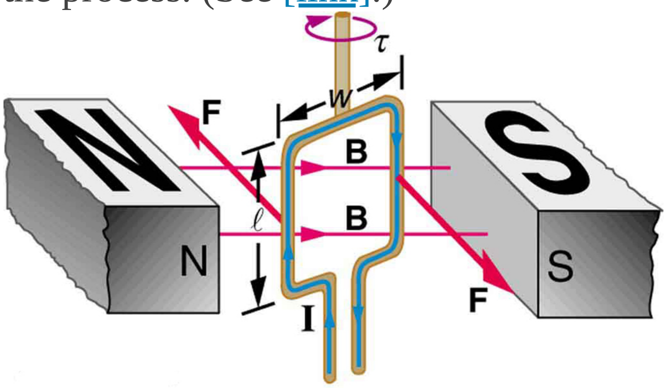


A rectangular loop of wire carrying a current is perpendicular to a magnetic field. The field is uniform in the region shown and is zero outside that region.

Torque on a Current Loop: Motors and Meters

- Describe how motors and meters work in terms of torque on a current loop.
- Calculate the torque on a current-carrying loop in a magnetic field.

Motors are the most common application of magnetic force on current-carrying wires. Motors have loops of wire in a magnetic field. When current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft. Electrical energy is converted to mechanical work in the process. (See [link].)



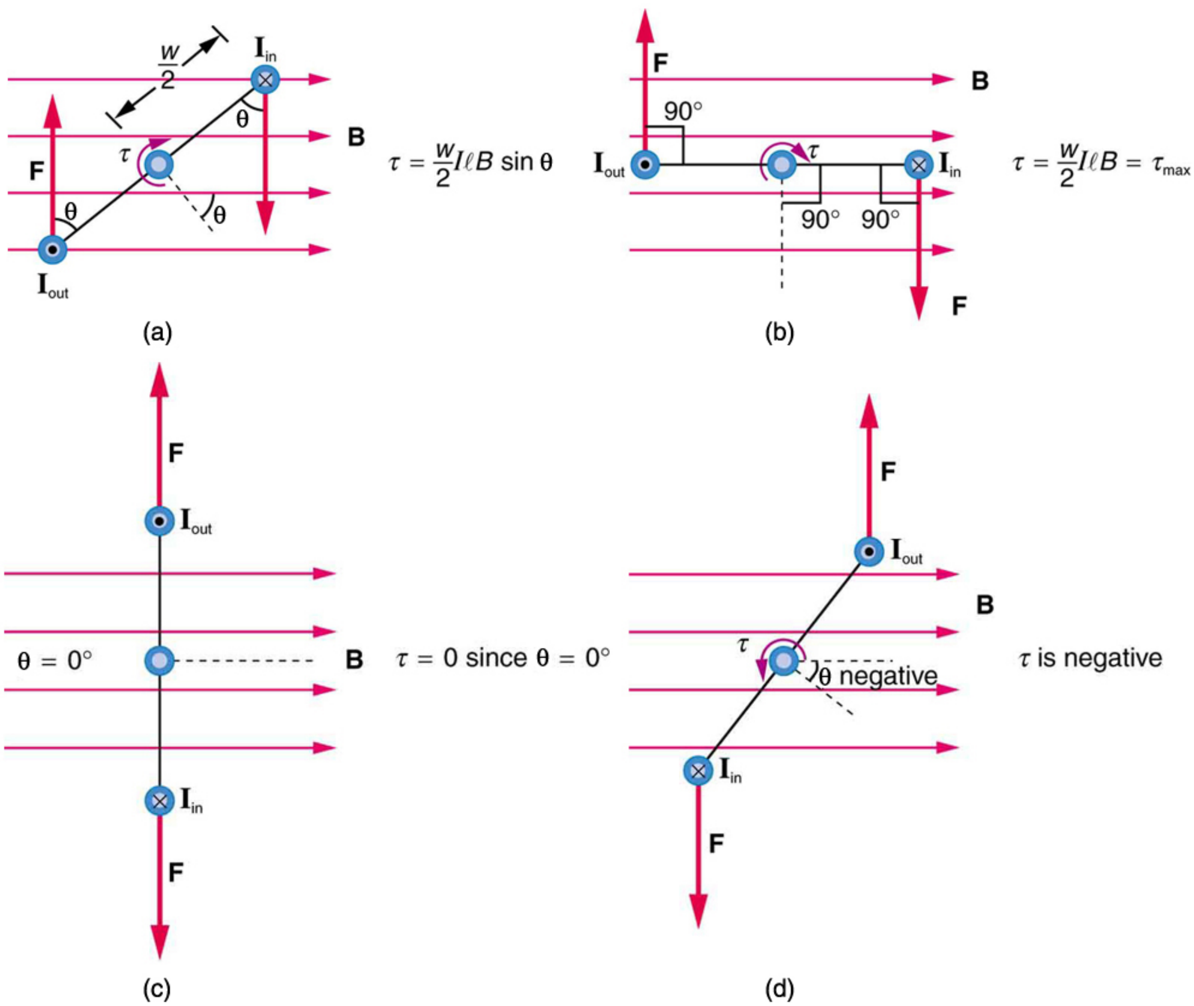
Torque on a current loop. A current-carrying loop of wire attached to a vertically rotating shaft feels magnetic forces that produce a clockwise torque as viewed from above.

Let us examine the force on each segment of the loop in [link] to find the torques produced about the axis of the vertical shaft. (This will lead to a useful equation for the torque on the loop.) We take the magnetic field to be uniform over the rectangular loop, which has width w and height l. First, we note that the forces on the top and bottom segments are vertical and, therefore, parallel to the shaft, producing no torque. Those vertical forces are equal in magnitude and opposite in direction, so that they also produce no net force on the loop. [link] shows views of the loop from above. Torque

is defined as $\tau = \mathrm{rF}\,\sin\,\theta$, where F is the force, r is the distance from the pivot that the force is applied, and θ is the angle between r and F. As seen in [link](a), right hand rule 1 gives the forces on the sides to be equal in magnitude and opposite in direction, so that the net force is again zero. However, each force produces a clockwise torque. Since $r = w/2$, the torque on each vertical segment is $(w/2)F\,\sin\,\theta$, and the two add to give a total torque.

Equation:

$$\tau = \frac{w}{2}F\,\sin\,\theta + \frac{w}{2}F\,\sin\,\theta = \mathrm{wF}\,\sin\,\theta$$



Top views of a current-carrying loop in a magnetic field. (a) The equation for torque is derived using this view. Note that the perpendicular to the loop makes an angle θ with the field that is the

same as the angle between $w/2$ and $\mathbf{F}$. (b) The maximum torque occurs when θ is a right angle and $\sin\,\theta = 1$. (c) Zero (minimum) torque occurs when θ is zero and $\sin\,\theta = 0$. (d) The torque reverses once the loop rotates past $\theta = 0$.

Now, each vertical segment has a length l that is perpendicular to B, so that the force on each is $F = \text{IlB}$. Entering F into the expression for torque yields

Equation:

$$\tau = \text{wIlB}\,\sin\,\theta.$$

If we have a multiple loop of N turns, we get N times the torque of one loop. Finally, note that the area of the loop is $A = \text{wl}$; the expression for the torque becomes

Equation:

$$\tau = \text{NIAB}\,\sin\,\theta.$$

This is the torque on a current-carrying loop in a uniform magnetic field. This equation can be shown to be valid for a loop of any shape. The loop carries a current I, has N turns, each of area A, and the perpendicular to the loop makes an angle θ with the field B. The net force on the loop is zero.

Example:
Calculating Torque on a Current-Carrying Loop in a Strong Magnetic Field

Find the maximum torque on a 100-turn square loop of a wire of 10.0 cm on a side that carries 15.0 A of current in a 2.00-T field.

Strategy

Torque on the loop can be found using $\tau = \text{NIAB}\,\sin\,\theta$. Maximum torque occurs when $\theta = 90^{\circ}$ and $\sin\,\theta = 1$.

Solution
For $\sin\theta = 1$, the maximum torque is
Equation:

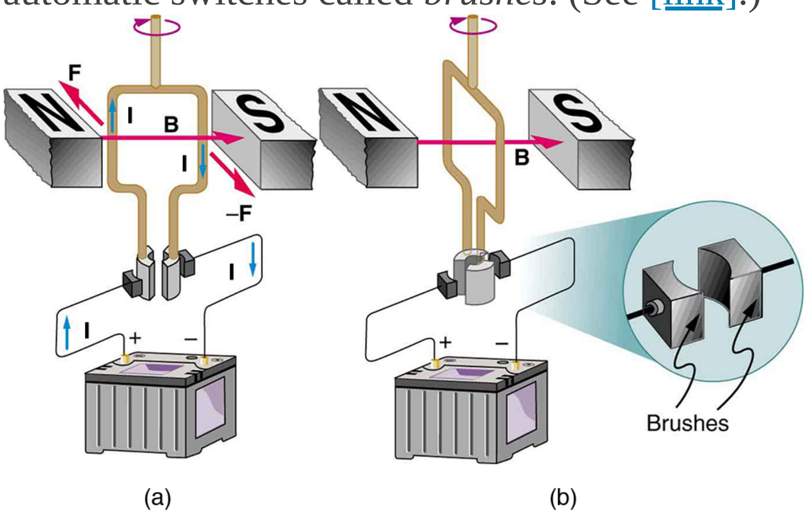
$$\tau_{\text{max}} = \text{NIAB}.$$

Entering known values yields
Equation:

$$\begin{array}{lll}\tau_{\text{max}} & = & (100)(15.0\text{ A})\left(0.100\text{ m}^2\right)(2.00\text{ T}) \\ & = & 30.0\text{ N}\cdot\text{m}.\end{array}$$

Discussion
This torque is large enough to be useful in a motor.

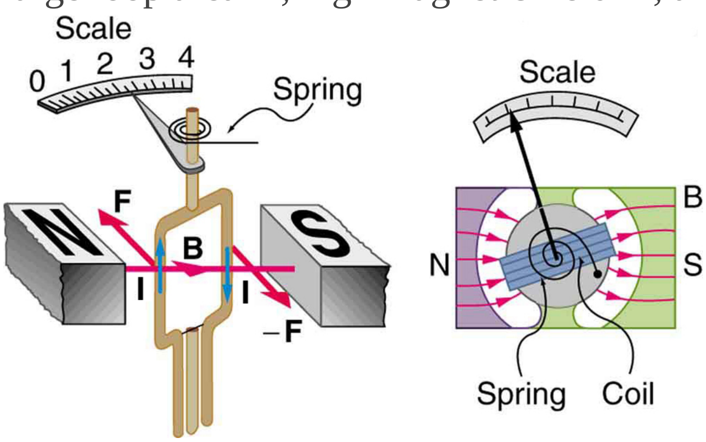
The torque found in the preceding example is the maximum. As the coil rotates, the torque decreases to zero at $\theta = 0$. The torque then *reverses* its direction once the coil rotates past $\theta = 0$. (See [link](d).) This means that, unless we do something, the coil will oscillate back and forth about equilibrium at $\theta = 0$. To get the coil to continue rotating in the same direction, we can reverse the current as it passes through $\theta = 0$ with automatic switches called *brushes*. (See [link].)



(a) As the angular momentum of the coil carries it through $\theta = 0$, the brushes reverse

the current to keep the torque clockwise. (b) The coil will rotate continuously in the clockwise direction, with the current reversing each half revolution to maintain the clockwise torque.

Meters, such as those in analog fuel gauges on a car, are another common application of magnetic torque on a current-carrying loop. [link] shows that a meter is very similar in construction to a motor. The meter in the figure has its magnets shaped to limit the effect of θ by making B perpendicular to the loop over a large angular range. Thus the torque is proportional to I and not θ. A linear spring exerts a counter-torque that balances the current-produced torque. This makes the needle deflection proportional to I. If an exact proportionality cannot be achieved, the gauge reading can be calibrated. To produce a galvanometer for use in analog voltmeters and ammeters that have a low resistance and respond to small currents, we use a large loop area A, high magnetic field B, and low-resistance coils.



Meters are very similar to motors but only rotate through a part of a revolution. The magnetic poles of this meter are shaped to keep the component of B perpendicular to the loop constant, so that the torque does not depend on θ and the deflection

against the return spring is proportional only to the current I.

Section Summary

- The torque τ on a current-carrying loop of any shape in a uniform magnetic field. is
  Equation:

  $$\tau = \text{NIAB}\sin\theta,$$

  where N is the number of turns, I is the current, A is the area of the loop, B is the magnetic field strength, and θ is the angle between the perpendicular to the loop and the magnetic field.

Conceptual Questions

Exercise:

Problem:

Draw a diagram and use RHR-1 to show that the forces on the top and bottom segments of the motor's current loop in [link] are vertical and produce no torque about the axis of rotation.

Problems & Exercises

Exercise:

Problem:

(a) By how many percent is the torque of a motor decreased if its permanent magnets lose 5.0% of their strength? (b) How many percent would the current need to be increased to return the torque to original values?

Solution:

(a) τ decreases by 5.00% if B decreases by 5.00%

(b) 5.26% increase

Exercise:

Problem:

(a) What is the maximum torque on a 150-turn square loop of wire 18.0 cm on a side that carries a 50.0-A current in a 1.60-T field? (b) What is the torque when θ is 10.9º?

Exercise:

Problem:

Find the current through a loop needed to create a maximum torque of $9.00\ \text{N}\cdot\text{m}$. The loop has 50 square turns that are 15.0 cm on a side and is in a uniform 0.800-T magnetic field.

Solution:

10.0 A

Exercise:

Problem:

Calculate the magnetic field strength needed on a 200-turn square loop 20.0 cm on a side to create a maximum torque of $300\ \text{N}\cdot\text{m}$ if the loop is carrying 25.0 A.

Exercise:

Problem:

Since the equation for torque on a current-carrying loop is $\tau = \text{NIAB}\sin\theta$, the units of $\text{N}\cdot\text{m}$ must equal units of $\text{A}\cdot\text{m}^2\ \text{T}$. Verify this.

Solution:

$$A \cdot m^2 \cdot T = A \cdot m^2 \left(\frac{N}{A \cdot m}\right) = N \cdot m.$$

Exercise:

Problem:

(a) At what angle θ is the torque on a current loop 90.0% of maximum? (b) 50.0% of maximum? (c) 10.0% of maximum?

Exercise:

Problem:

A proton has a magnetic field due to its spin on its axis. The field is similar to that created by a circular current loop $0.650 \times 10^{-15}\ \text{m}$ in radius with a current of $1.05 \times 10^4\ \text{A}$ (no kidding). Find the maximum torque on a proton in a 2.50-T field. (This is a significant torque on a small particle.)

Solution:

$3.48 \times 10^{-26}\ \text{N} \cdot \text{m}$

Exercise:

Problem:

(a) A 200-turn circular loop of radius 50.0 cm is vertical, with its axis on an east-west line. A current of 100 A circulates clockwise in the loop when viewed from the east. The Earth's field here is due north, parallel to the ground, with a strength of $3.00 \times 10^{-5}\ \text{T}$. What are the direction and magnitude of the torque on the loop? (b) Does this device have any practical applications as a motor?

Exercise:

Problem:

Repeat [link], but with the loop lying flat on the ground with its current circulating counterclockwise (when viewed from above) in a location where the Earth's field is north, but at an angle 45.0^{o} below the horizontal and with a strength of $6.00 \times 10^{-5}\ \text{T}$.

Solution:

(a) $0.666\ \text{N}\cdot\text{m}$ west

(b) This is not a very significant torque, so practical use would be limited. Also, the current would need to be alternated to make the loop rotate (otherwise it would oscillate).

Glossary

motor
: loop of wire in a magnetic field; when current is passed through the loops, the magnetic field exerts torque on the loops, which rotates a shaft; electrical energy is converted to mechanical work in the process

meter
: common application of magnetic torque on a current-carrying loop that is very similar in construction to a motor; by design, the torque is proportional to I and not θ, so the needle deflection is proportional to the current

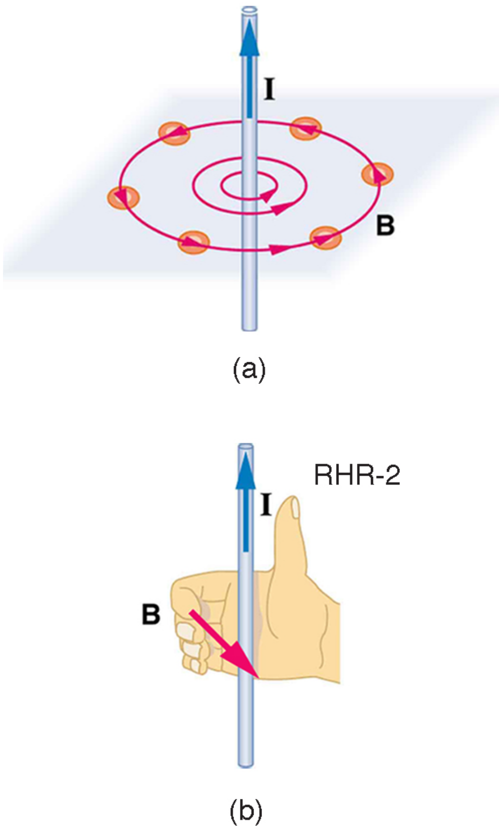
Magnetic Fields Produced by Currents: Ampere’s Law

- Calculate current that produces a magnetic field.
- Use the right hand rule 2 to determine the direction of current or the direction of magnetic field loops.

How much current is needed to produce a significant magnetic field, perhaps as strong as the Earth’s field? Surveyors will tell you that overhead electric power lines create magnetic fields that interfere with their compass readings. Indeed, when Oersted discovered in 1820 that a current in a wire affected a compass needle, he was not dealing with extremely large currents. How does the shape of wires carrying current affect the shape of the magnetic field created? We noted earlier that a current loop created a magnetic field similar to that of a bar magnet, but what about a straight wire or a toroid (doughnut)? How is the direction of a current-created field related to the direction of the current? Answers to these questions are explored in this section, together with a brief discussion of the law governing the fields created by currents.

Magnetic Field Created by a Long Straight Current-Carrying Wire: Right Hand Rule 2

Magnetic fields have both direction and magnitude. As noted before, one way to explore the direction of a magnetic field is with compasses, as shown for a long straight current-carrying wire in [link]. Hall probes can determine the magnitude of the field. The field around a long straight wire is found to be in circular loops. The **right hand rule 2** (RHR-2) emerges from this exploration and is valid for any current segment—*point the thumb in the direction of the current, and the fingers curl in the direction of the magnetic field loops* created by it.



(a) Compasses placed near a long straight current-carrying wire indicate that field lines form circular loops centered on the wire. (b) Right hand rule 2 states that, if the right hand thumb points in the direction of the current, the fingers curl in the direction of the field. This rule is consistent with the field mapped for the long straight wire and is valid for any current segment.

The **magnetic field strength (magnitude) produced by a long straight current-carrying wire** is found by experiment to be

Equation:

$$B = \frac{\mu_0 I}{2\pi r} \text{ (long straight wire)},$$

where I is the current, r is the shortest distance to the wire, and the constant $\mu_0 = 4\pi \times 10^{-7}\ \text{T} \cdot \text{m/A}$ is the **permeability of free space**. (μ_0 is one of the basic constants in nature. We will see later that μ_0 is related to the speed of light.) Since the wire is very long, the magnitude of the field depends only on distance from the wire r, not on position along the wire.

Example:

Calculating Current that Produces a Magnetic Field

Find the current in a long straight wire that would produce a magnetic field twice the strength of the Earth's at a distance of 5.0 cm from the wire.

Strategy

The Earth's field is about 5.0×10^{-5} T, and so here B due to the wire is taken to be 1.0×10^{-4} T. The equation $B = \frac{\mu_0 I}{2\pi r}$ can be used to find I, since all other quantities are known.

Solution

Solving for I and entering known values gives

Equation:

$$\begin{aligned} I &= \frac{2\pi r B}{\mu_0} = \frac{2\pi\left(5.0\times10^{-2}\text{ m}\right)\left(1.0\times10^{-4}\text{ T}\right)}{4\pi\times10^{-7}\text{ T}\cdot\text{m/A}} \\ &= 25\text{ A}. \end{aligned}$$

Discussion

So a moderately large current produces a significant magnetic field at a distance of 5.0 cm from a long straight wire. Note that the answer is stated to only two digits, since the Earth's field is specified to only two digits in this example.

Ampere’s Law and Others

The magnetic field of a long straight wire has more implications than you might at first suspect. *Each segment of current produces a magnetic field like that of a long straight wire, and the total field of any shape current is the vector sum of the fields due to each segment.* The formal statement of the direction and magnitude of the field due to each segment is called the **Biot-Savart law**. Integral calculus is needed to sum the field for an arbitrary shape current. This results in a more complete law, called **Ampere’s law**, which relates magnetic field and current in a general way. Ampere’s law in turn is a part of **Maxwell’s equations**, which give a complete theory of all electromagnetic phenomena. Considerations of how Maxwell’s equations appear to different observers led to the modern theory of relativity, and the realization that electric and magnetic fields are different manifestations of the same thing. Most of this is beyond the scope of this text in both mathematical level, requiring calculus, and in the amount of space that can be devoted to it. But for the interested student, and particularly for those who continue in physics, engineering, or similar pursuits, delving into these matters further will reveal descriptions of nature that are elegant as well as profound. In this text, we shall keep the general features in mind, such as RHR-2 and the rules for magnetic field lines listed in Magnetic Fields and Magnetic Field Lines, while concentrating on the fields created in certain important situations.

> **Note:**
> Making Connections: Relativity
> Hearing all we do about Einstein, we sometimes get the impression that he invented relativity out of nothing. On the contrary, one of Einstein’s motivations was to solve difficulties in knowing how different observers see magnetic and electric fields.

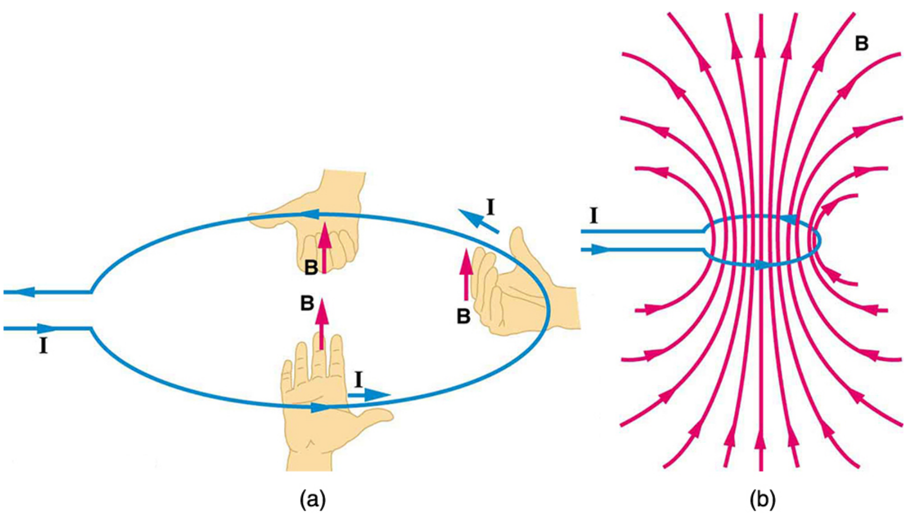
Magnetic Field Produced by a Current-Carrying Circular Loop

The magnetic field near a current-carrying loop of wire is shown in [link]. Both the direction and the magnitude of the magnetic field produced by a current-carrying loop are complex. RHR-2 can be used to give the direction of the field near the loop, but mapping with compasses and the rules about field lines given in Magnetic Fields and Magnetic Field Lines are needed for more detail. There is a simple formula for the **magnetic field strength at the center of a circular loop**. It is

Equation:

$$B = \frac{\mu_0 I}{2R} \text{ (at center of loop)},$$

where R is the radius of the loop. This equation is very similar to that for a straight wire, but it is valid *only* at the center of a circular loop of wire. The similarity of the equations does indicate that similar field strength can be obtained at the center of a loop. One way to get a larger field is to have N loops; then, the field is $B = N\mu_0 I/(2R)$. Note that the larger the loop, the smaller the field at its center, because the current is farther away.

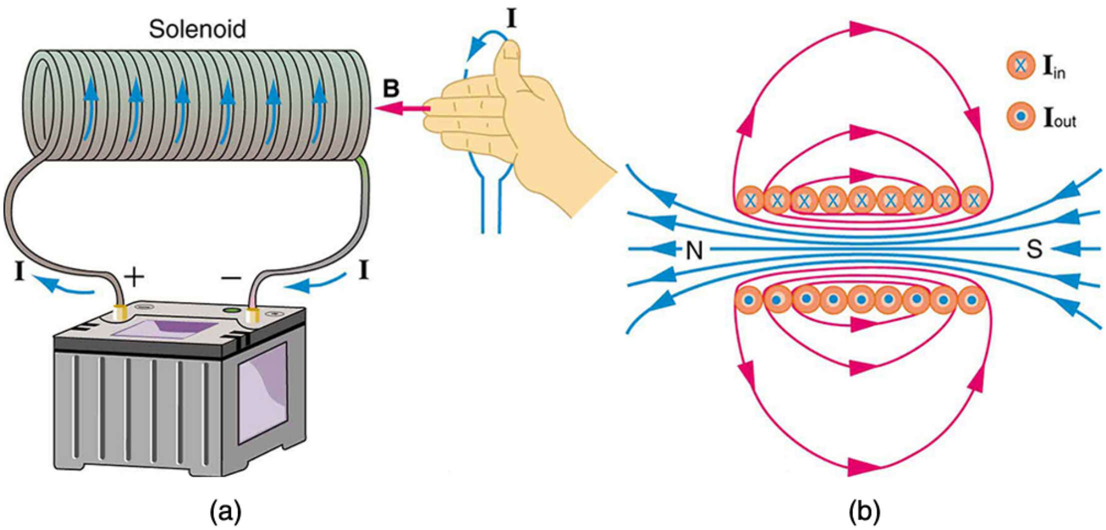


(a) RHR-2 gives the direction of the magnetic field inside and outside a current-carrying loop. (b) More detailed mapping with compasses or with a

Hall probe completes the picture. The field is similar to that of a bar magnet.

Magnetic Field Produced by a Current-Carrying Solenoid

A **solenoid** is a long coil of wire (with many turns or loops, as opposed to a flat loop). Because of its shape, the field inside a solenoid can be very uniform, and also very strong. The field just outside the coils is nearly zero. [link] shows how the field looks and how its direction is given by RHR-2.



(a) Because of its shape, the field inside a solenoid of length l is remarkably uniform in magnitude and direction, as indicated by the straight and uniformly spaced field lines. The field outside the coils is nearly zero. (b) This cutaway shows the magnetic field generated by the current in the solenoid.

The magnetic field inside of a current-carrying solenoid is very uniform in direction and magnitude. Only near the ends does it begin to weaken and change direction. The field outside has similar complexities to flat loops

and bar magnets, but the **magnetic field strength inside a solenoid** is simply

Equation:

$$B = \mu_0 \text{nI} \ \ (\text{inside a solenoid}),$$

where n is the number of loops per unit length of the solenoid $(n = N/l$, with N being the number of loops and l the length). Note that B is the field strength anywhere in the uniform region of the interior and not just at the center. Large uniform fields spread over a large volume are possible with solenoids, as [link] implies.

Example:

Calculating Field Strength inside a Solenoid

What is the field inside a 2.00-m-long solenoid that has 2000 loops and carries a 1600-A current?

Strategy

To find the field strength inside a solenoid, we use $B = \mu_0 \text{nI}$. First, we note the number of loops per unit length is

Equation:

$$n = \frac{N}{l} = \frac{2000}{2.00 \text{ m}} = 1000 \text{ m}^{-1} = 10 \text{ cm}^{-1}.$$

Solution

Substituting known values gives

Equation:

$$\begin{aligned} B &= \mu_0 \text{nI} = \left(4\pi \times 10^{-7} \text{ T} \cdot \text{m/A}\right)\left(1000 \text{ m}^{-1}\right)\left(1600 \text{ A}\right) \\ &= 2.01 \text{ T}. \end{aligned}$$

Discussion

This is a large field strength that could be established over a large-diameter solenoid, such as in medical uses of magnetic resonance imaging (MRI). The very large current is an indication that the fields of this strength are not

easily achieved, however. Such a large current through 1000 loops squeezed into a meter’s length would produce significant heating. Higher currents can be achieved by using superconducting wires, although this is expensive. There is an upper limit to the current, since the superconducting state is disrupted by very large magnetic fields.

There are interesting variations of the flat coil and solenoid. For example, the toroidal coil used to confine the reactive particles in tokamaks is much like a solenoid bent into a circle. The field inside a toroid is very strong but circular. Charged particles travel in circles, following the field lines, and collide with one another, perhaps inducing fusion. But the charged particles do not cross field lines and escape the toroid. A whole range of coil shapes are used to produce all sorts of magnetic field shapes. Adding ferromagnetic materials produces greater field strengths and can have a significant effect on the shape of the field. Ferromagnetic materials tend to trap magnetic fields (the field lines bend into the ferromagnetic material, leaving weaker fields outside it) and are used as shields for devices that are adversely affected by magnetic fields, including the Earth’s magnetic field.

Note:
PhET Explorations: Generator
Generate electricity with a bar magnet! Discover the physics behind the phenomena by exploring magnets and how you can use them to make a bulb light.

Generator

Section Summary

- The strength of the magnetic field created by current in a long straight wire is given by
  Equation:

  $$B = \frac{\mu_0 I}{2\pi r} (\text{long straight wire}),$$

  where I is the current, r is the shortest distance to the wire, and the constant $\mu_0 = 4\pi \times 10^{-7} \text{ T} \cdot \text{m/A}$ is the permeability of free space.
- The direction of the magnetic field created by a long straight wire is given by right hand rule 2 (RHR-2): *Point the thumb of the right hand in the direction of current, and the fingers curl in the direction of the magnetic field loops* created by it.
- The magnetic field created by current following any path is the sum (or integral) of the fields due to segments along the path (magnitude and direction as for a straight wire), resulting in a general relationship between current and field known as Ampere's law.
- The magnetic field strength at the center of a circular loop is given by
  Equation:

  $$B = \frac{\mu_0 I}{2R} (\text{at center of loop}),$$

  where R is the radius of the loop. This equation becomes $B = \mu_0 \text{nI}/(2R)$ for a flat coil of N loops. RHR-2 gives the direction of the field about the loop. A long coil is called a solenoid.
- The magnetic field strength inside a solenoid is
  Equation:

  $$B = \mu_0 \text{nI} \ \ (\text{inside a solenoid}),$$

  where n is the number of loops per unit length of the solenoid. The field inside is very uniform in magnitude and direction.

Conceptual Questions

Exercise:

Problem:

Make a drawing and use RHR-2 to find the direction of the magnetic field of a current loop in a motor (such as in [link]). Then show that the direction of the torque on the loop is the same as produced by like poles repelling and unlike poles attracting.

Glossary

right hand rule 2 (RHR-2)
: a rule to determine the direction of the magnetic field induced by a current-carrying wire: Point the thumb of the right hand in the direction of current, and the fingers curl in the direction of the magnetic field loops

magnetic field strength (magnitude) produced by a long straight current-carrying wire
: defined as $B = \frac{\mu_0 I}{2\pi r}$, where I is the current, r is the shortest distance to the wire, and μ_0 is the permeability of free space

permeability of free space
: the measure of the ability of a material, in this case free space, to support a magnetic field; the constant $\mu_0 = 4\pi \times 10^{-7}\ \mathrm{T \cdot m/A}$

magnetic field strength at the center of a circular loop
: defined as $B = \frac{\mu_0 I}{2R}$ where R is the radius of the loop

solenoid
: a thin wire wound into a coil that produces a magnetic field when an electric current is passed through it

magnetic field strength inside a solenoid
: defined as $B = \mu_0 n I$ where n is the number of loops per unit length of the solenoid ($n = N/l$, with N being the number of loops and l the length)

Biot-Savart law

a physical law that describes the magnetic field generated by an electric current in terms of a specific equation

Ampere’s law

the physical law that states that the magnetic field around an electric current is proportional to the current; each segment of current produces a magnetic field like that of a long straight wire, and the total field of any shape current is the vector sum of the fields due to each segment

Maxwell’s equations

a set of four equations that describe electromagnetic phenomena

Magnetic Force between Two Parallel Conductors

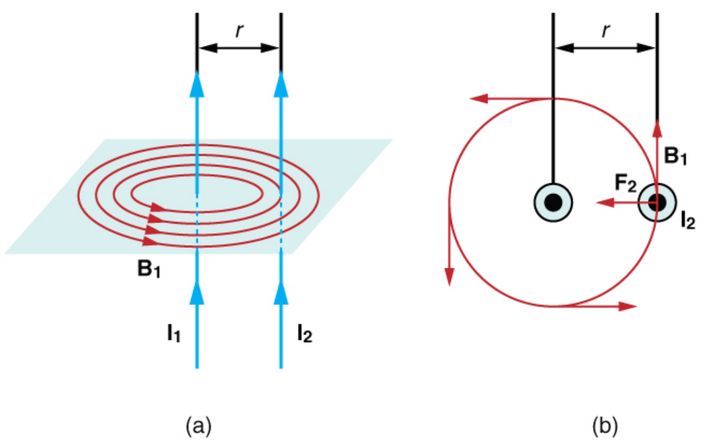
- Describe the effects of the magnetic force between two conductors.
- Calculate the force between two parallel conductors.

You might expect that there are significant forces between current-carrying wires, since ordinary currents produce significant magnetic fields and these fields exert significant forces on ordinary currents. But you might not expect that the force between wires is used to *define* the ampere. It might also surprise you to learn that this force has something to do with why large circuit breakers burn up when they attempt to interrupt large currents.

The force between two long straight and parallel conductors separated by a distance r can be found by applying what we have developed in preceding sections. [link] shows the wires, their currents, the fields they create, and the subsequent forces they exert on one another. Let us consider the field produced by wire 1 and the force it exerts on wire 2 (call the force F_2). The field due to I_1 at a distance r is given to be

Equation:

$$B_1 = \frac{\mu_0 I_1}{2\pi r}.$$



(a) The magnetic field produced by a long straight conductor is perpendicular to a parallel conductor, as indicated by RHR-2. (b) A view

from above of the two wires shown in (a), with one magnetic field line shown for each wire. RHR-1 shows that the force between the parallel conductors is attractive when the currents are in the same direction. A similar analysis shows that the force is repulsive between currents in opposite directions.

This field is uniform along wire 2 and perpendicular to it, and so the force F_2 it exerts on wire 2 is given by $F = \text{IlB} \sin \theta$ with $\sin \theta = 1$:

Equation:

$$F_2 = I_2 l B_1.$$

By Newton's third law, the forces on the wires are equal in magnitude, and so we just write F for the magnitude of F_2. (Note that $F_1 = -F_2$.) Since the wires are very long, it is convenient to think in terms of F/l, the force per unit length. Substituting the expression for B_1 into the last equation and rearranging terms gives

Equation:

$$\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}.$$

F/l is the force per unit length between two parallel currents I_1 and I_2 separated by a distance r. The force is attractive if the currents are in the same direction and repulsive if they are in opposite directions.

This force is responsible for the *pinch effect* in electric arcs and plasmas. The force exists whether the currents are in wires or not. In an electric arc, where currents are moving parallel to one another, there is an attraction that squeezes currents into a smaller tube. In large circuit breakers, like those

used in neighborhood power distribution systems, the pinch effect can concentrate an arc between plates of a switch trying to break a large current, burn holes, and even ignite the equipment. Another example of the pinch effect is found in the solar plasma, where jets of ionized material, such as solar flares, are shaped by magnetic forces.

The *operational definition of the ampere* is based on the force between current-carrying wires. Note that for parallel wires separated by 1 meter with each carrying 1 ampere, the force per meter is

Equation:

$$\frac{F}{l} = \frac{\left(4\pi \times 10^{-7}\ \text{T} \cdot \text{m/A}\right)(1\ \text{A})^2}{(2\pi)(1\ \text{m})} = 2 \times 10^{-7}\ \text{N/m}.$$

Since μ_0 is exactly $4\pi \times 10^{-7}\ \text{T} \cdot \text{m/A}$ by definition, and because $1\ \text{T} = 1\ \text{N/(A} \cdot \text{m)}$, the force per meter is exactly $2 \times 10^{-7}\ \text{N/m}$. This is the basis of the operational definition of the ampere.

Note:
The Ampere
The official definition of the ampere is:
One ampere of current through each of two parallel conductors of infinite length, separated by one meter in empty space free of other magnetic fields, causes a force of exactly $2 \times 10^{-7}\ \text{N/m}$ on each conductor.

Infinite-length straight wires are impractical and so, in practice, a current balance is constructed with coils of wire separated by a few centimeters. Force is measured to determine current. This also provides us with a method for measuring the coulomb. We measure the charge that flows for a current of one ampere in one second. That is, $1\ \text{C} = 1\ \text{A} \cdot \text{s}$. For both the ampere and the coulomb, the method of measuring force between conductors is the most accurate in practice.

Section Summary

- The force between two parallel currents I_1 and I_2, separated by a distance r, has a magnitude per unit length given by
  Equation:

  $$\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}.$$

- The force is attractive if the currents are in the same direction, repulsive if they are in opposite directions.

Conceptual Questions

Exercise:

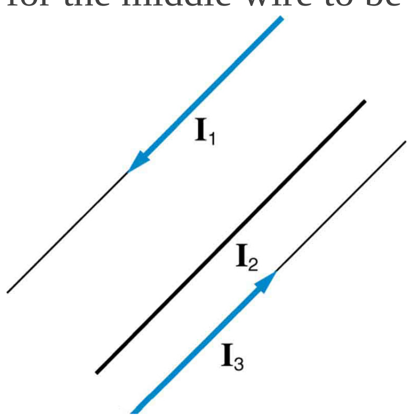
Problem:

Is the force attractive or repulsive between the hot and neutral lines hung from power poles? Why?

Exercise:

Problem:

If you have three parallel wires in the same plane, as in [link], with currents in the outer two running in opposite directions, is it possible for the middle wire to be repelled by both? Attracted by both? Explain.



Three parallel

coplanar wires with currents in the outer two in opposite directions.

Exercise:

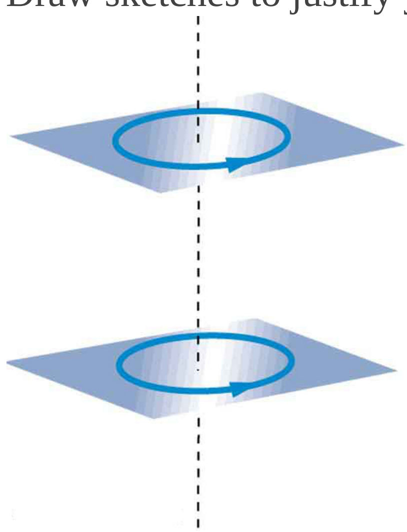
Problem:

Suppose two long straight wires run perpendicular to one another without touching. Does one exert a net force on the other? If so, what is its direction? Does one exert a net torque on the other? If so, what is its direction? Justify your responses by using the right hand rules.

Exercise:

Problem:

Use the right hand rules to show that the force between the two loops in [link] is attractive if the currents are in the same direction and repulsive if they are in opposite directions. Is this consistent with like poles of the loops repelling and unlike poles of the loops attracting? Draw sketches to justify your answers.



Two loops of wire carrying currents

can exert forces and torques on one another.

Exercise:

Problem:

If one of the loops in [link] is tilted slightly relative to the other and their currents are in the same direction, what are the directions of the torques they exert on each other? Does this imply that the poles of the bar magnet-like fields they create will line up with each other if the loops are allowed to rotate?

Exercise:

Problem:

Electric field lines can be shielded by the Faraday cage effect. Can we have magnetic shielding? Can we have gravitational shielding?

Problems & Exercises

Exercise:

Problem:

(a) The hot and neutral wires supplying DC power to a light-rail commuter train carry 800 A and are separated by 75.0 cm. What is the magnitude and direction of the force between 50.0 m of these wires? (b) Discuss the practical consequences of this force, if any.

Solution:

(a) 8.53 N, repulsive

(b) This force is repulsive and therefore there is never a risk that the two wires will touch and short circuit.

Exercise:

Problem:

The force per meter between the two wires of a jumper cable being used to start a stalled car is 0.225 N/m. (a) What is the current in the wires, given they are separated by 2.00 cm? (b) Is the force attractive or repulsive?

Exercise:

Problem:

A 2.50-m segment of wire supplying current to the motor of a submerged submarine carries 1000 A and feels a 4.00-N repulsive force from a parallel wire 5.00 cm away. What is the direction and magnitude of the current in the other wire?

Solution:

400 A in the opposite direction

Exercise:

Problem:

The wire carrying 400 A to the motor of a commuter train feels an attractive force of $4.00 \times 10^{-3}\ \mathrm{N/m}$ due to a parallel wire carrying 5.00 A to a headlight. (a) How far apart are the wires? (b) Are the currents in the same direction?

Exercise:

Problem:

An AC appliance cord has its hot and neutral wires separated by 3.00 mm and carries a 5.00-A current. (a) What is the average force per meter between the wires in the cord? (b) What is the maximum force per meter between the wires? (c) Are the forces attractive or repulsive? (d) Do appliance cords need any special design features to compensate for these forces?

Solution:

(a) $1.67 \times 10^{-3}\ \text{N/m}$

(b) $3.33 \times 10^{-3}\ \text{N/m}$

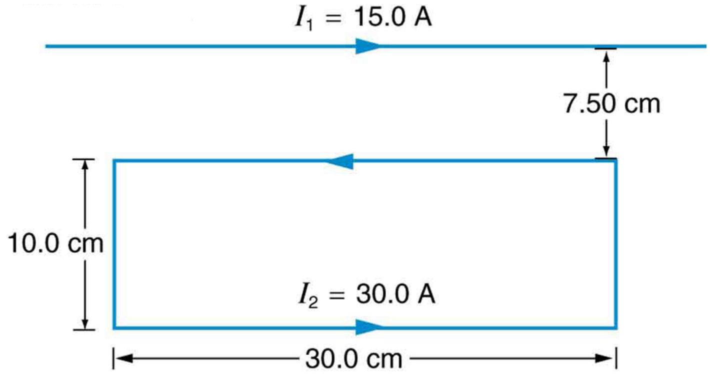
(c) Repulsive

(d) No, these are very small forces

Exercise:

Problem:

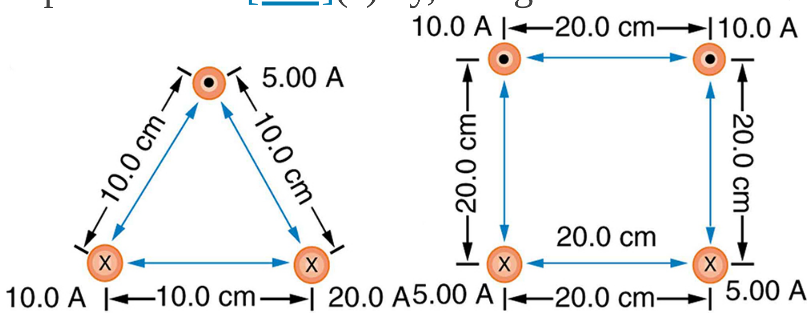
[link] shows a long straight wire near a rectangular current loop. What is the direction and magnitude of the total force on the loop?



Exercise:

Problem:

Find the direction and magnitude of the force that each wire experiences in [link](a) by, using vector addition.



Solution:

(a) Top wire: $2.65 \times 10^{-4}\ \mathrm{N/m}$ s, 10.9^o to left of up

(b) Lower left wire: $3.61 \times 10^{-4}\ \mathrm{N/m}$, 13.9^o down from right

(c) Lower right wire: $3.46 \times 10^{-4}\ \mathrm{N/m}$, 30.0^o down from left

Exercise:

Problem:

Find the direction and magnitude of the force that each wire experiences in [link](b), using vector addition.

More Applications of Magnetism

- Describe some applications of magnetism.

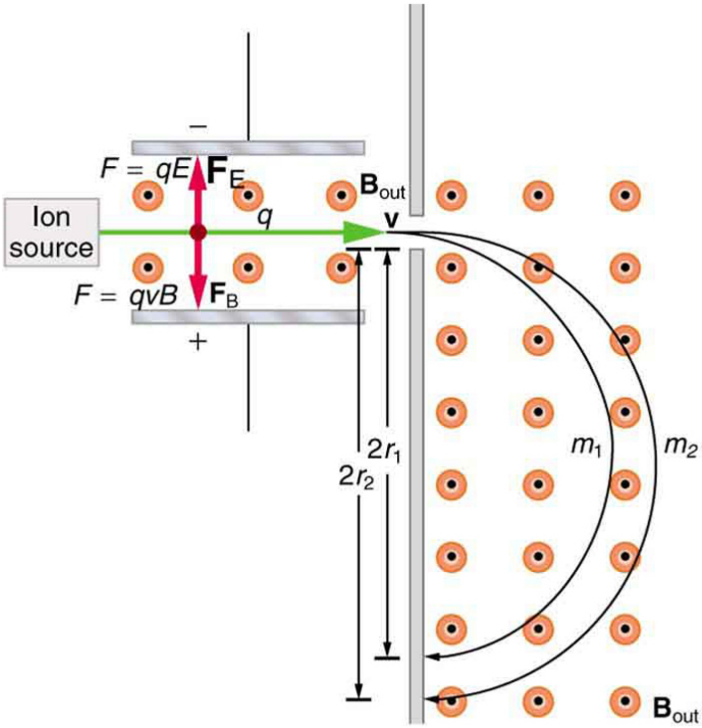
Mass Spectrometry

The curved paths followed by charged particles in magnetic fields can be put to use. A charged particle moving perpendicular to a magnetic field travels in a circular path having a radius r.

Equation:

$$r = \frac{\text{mv}}{\text{qB}}$$

It was noted that this relationship could be used to measure the mass of charged particles such as ions. A mass spectrometer is a device that measures such masses. Most mass spectrometers use magnetic fields for this purpose, although some of them have extremely sophisticated designs. Since there are five variables in the relationship, there are many possibilities. However, if v, q, and B can be fixed, then the radius of the path r is simply proportional to the mass m of the charged particle. Let us examine one such mass spectrometer that has a relatively simple design. (See [link].) The process begins with an ion source, a device like an electron gun. The ion source gives ions their charge, accelerates them to some velocity v, and directs a beam of them into the next stage of the spectrometer. This next region is a *velocity selector* that only allows particles with a particular value of v to get through.



This mass spectrometer uses a velocity selector to fix v so that the radius of the path is proportional to mass.

The velocity selector has both an electric field and a magnetic field, perpendicular to one another, producing forces in opposite directions on the ions. Only those ions for which the forces balance travel in a straight line into the next region. If the forces balance, then the electric force $F = \text{qE}$ equals the magnetic force $F = \text{qvB}$, so that $\text{qE} = \text{qvB}$. Noting that q cancels, we see that

Equation:

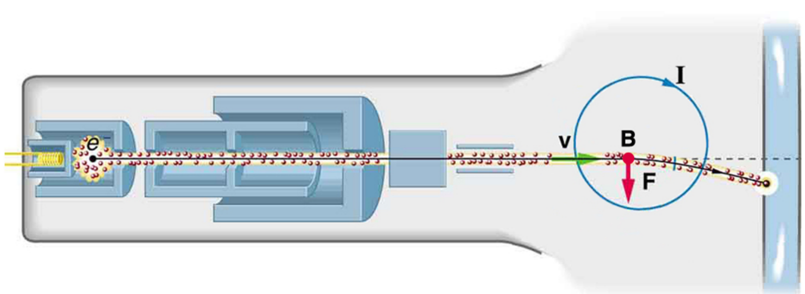
$$v = \frac{E}{B}$$

is the velocity particles must have to make it through the velocity selector, and further, that v can be selected by varying E and B. In the final region, there is only a uniform magnetic field, and so the charged particles move in circular arcs with radii proportional to particle mass. The paths also depend on charge q, but since q is in multiples of electron charges, it is easy to determine and to discriminate between ions in different charge states.

Mass spectrometry today is used extensively in chemistry and biology laboratories to identify chemical and biological substances according to their mass-to-charge ratios. In medicine, mass spectrometers are used to measure the concentration of isotopes used as tracers. Usually, biological molecules such as proteins are very large, so they are broken down into smaller fragments before analyzing. Recently, large virus particles have been analyzed as a whole on mass spectrometers. Sometimes a gas chromatograph or high-performance liquid chromatograph provides an initial separation of the large molecules, which are then input into the mass spectrometer.

Cathode Ray Tubes—CRTs—and the Like

What do non-flat-screen TVs, old computer monitors, x-ray machines, and the 2-mile-long Stanford Linear Accelerator have in common? All of them accelerate electrons, making them different versions of the electron gun. Many of these devices use magnetic fields to steer the accelerated electrons. [link] shows the construction of the type of cathode ray tube (CRT) found in some TVs, oscilloscopes, and old computer monitors. Two pairs of coils are used to steer the electrons, one vertically and the other horizontally, to their desired destination.



The cathode ray tube (CRT) is so named because rays of electrons originate at the cathode in the electron gun. Magnetic coils are used to steer the beam in many CRTs. In this case, the beam is moved down. Another pair of horizontal coils would steer the beam horizontally.

Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is one of the most useful and rapidly growing medical imaging tools. It non-invasively produces two-dimensional and three-dimensional images of the body that provide important medical information with none of the hazards of x-rays. MRI is based on an effect called **nuclear magnetic resonance (NMR)** in which an externally applied magnetic field interacts with the nuclei of certain atoms, particularly those of hydrogen (protons). These nuclei possess their own small magnetic fields, similar to those of electrons and the current loops discussed earlier in this chapter.

When placed in an external magnetic field, such nuclei experience a torque that pushes or aligns the nuclei into one of two new energy states—depending on the orientation of its spin (analogous to the N pole and S pole in a bar magnet). Transitions from the lower to higher energy state can be achieved by using an external radio frequency signal to “flip” the orientation of the small magnets. (This is actually a quantum mechanical process. The direction of the nuclear magnetic field is quantized as is energy in the radio waves. We will return to these topics in later chapters.) The specific frequency of the radio waves that are absorbed and reemitted depends sensitively on the type of nucleus, the chemical environment, and the external magnetic field strength. Therefore, this is a *resonance* phenomenon in which *nuclei* in a *magnetic* field act like resonators (analogous to those discussed in the treatment of sound in Oscillatory

Motion and Waves) that absorb and reemit only certain frequencies. Hence, the phenomenon is named *nuclear magnetic resonance (NMR).*

NMR has been used for more than 50 years as an analytical tool. It was formulated in 1946 by F. Bloch and E. Purcell, with the 1952 Nobel Prize in Physics going to them for their work. Over the past two decades, NMR has been developed to produce detailed images in a process now called magnetic resonance imaging (MRI), a name coined to avoid the use of the word “nuclear” and the concomitant implication that nuclear radiation is involved. (It is not.) The 2003 Nobel Prize in Medicine went to P. Lauterbur and P. Mansfield for their work with MRI applications.

The largest part of the MRI unit is a superconducting magnet that creates a magnetic field, typically between 1 and 2 T in strength, over a relatively large volume. MRI images can be both highly detailed and informative about structures and organ functions. It is helpful that normal and non-normal tissues respond differently for slight changes in the magnetic field. In most medical images, the protons that are hydrogen nuclei are imaged. (About 2/3 of the atoms in the body are hydrogen.) Their location and density give a variety of medically useful information, such as organ function, the condition of tissue (as in the brain), and the shape of structures, such as vertebral disks and knee-joint surfaces. MRI can also be used to follow the movement of certain ions across membranes, yielding information on active transport, osmosis, dialysis, and other phenomena. With excellent spatial resolution, MRI can provide information about tumors, strokes, shoulder injuries, infections, etc.

An image requires position information as well as the density of a nuclear type (usually protons). By varying the magnetic field slightly over the volume to be imaged, the resonant frequency of the protons is made to vary with position. Broadcast radio frequencies are swept over an appropriate range and nuclei absorb and reemit them only if the nuclei are in a magnetic field with the correct strength. The imaging receiver gathers information through the body almost point by point, building up a tissue map. The reception of reemitted radio waves as a function of frequency thus gives position information. These “slices” or cross sections through the body are only several mm thick. The intensity of the reemitted radio waves is

proportional to the concentration of the nuclear type being flipped, as well as information on the chemical environment in that area of the body. Various techniques are available for enhancing contrast in images and for obtaining more information. Scans called T1, T2, or proton density scans rely on different relaxation mechanisms of nuclei. Relaxation refers to the time it takes for the protons to return to equilibrium after the external field is turned off. This time depends upon tissue type and status (such as inflammation).

While MRI images are superior to x rays for certain types of tissue and have none of the hazards of x rays, they do not completely supplant x-ray images. MRI is less effective than x rays for detecting breaks in bone, for example, and in imaging breast tissue, so the two diagnostic tools complement each other. MRI images are also expensive compared to simple x-ray images and tend to be used most often where they supply information not readily obtained from x rays. Another disadvantage of MRI is that the patient is totally enclosed with detectors close to the body for about 30 minutes or more, leading to claustrophobia. It is also difficult for the obese patient to be in the magnet tunnel. New “open-MRI” machines are now available in which the magnet does not completely surround the patient.

Over the last decade, the development of much faster scans, called “functional MRI” (fMRI), has allowed us to map the functioning of various regions in the brain responsible for thought and motor control. This technique measures the change in blood flow for activities (thought, experiences, action) in the brain. The nerve cells increase their consumption of oxygen when active. Blood hemoglobin releases oxygen to active nerve cells and has somewhat different magnetic properties when oxygenated than when deoxygenated. With MRI, we can measure this and detect a blood oxygen-dependent signal. Most of the brain scans today use fMRI.

Other Medical Uses of Magnetic Fields

Currents in nerve cells and the heart create magnetic fields like any other currents. These can be measured but with some difficulty since their strengths are about 10^{-6} to 10^{-8} *less* than the Earth’s magnetic field. Recording of the heart’s magnetic field as it beats is called a

magnetocardiogram (MCG), while measurements of the brain’s magnetic field is called a **magnetoencephalogram (MEG)**. Both give information that differs from that obtained by measuring the electric fields of these organs (ECGs and EEGs), but they are not yet of sufficient importance to make these difficult measurements common.

In both of these techniques, the sensors do not touch the body. MCG can be used in fetal studies, and is probably more sensitive than echocardiography. MCG also looks at the heart’s electrical activity whose voltage output is too small to be recorded by surface electrodes as in EKG. It has the potential of being a rapid scan for early diagnosis of cardiac ischemia (obstruction of blood flow to the heart) or problems with the fetus.

MEG can be used to identify abnormal electrical discharges in the brain that produce weak magnetic signals. Therefore, it looks at brain activity, not just brain structure. It has been used for studies of Alzheimer’s disease and epilepsy. Advances in instrumentation to measure very small magnetic fields have allowed these two techniques to be used more in recent years. What is used is a sensor called a SQUID, for superconducting quantum interference device. This operates at liquid helium temperatures and can measure magnetic fields thousands of times smaller than the Earth’s.

Finally, there is a burgeoning market for magnetic cures in which magnets are applied in a variety of ways to the body, from magnetic bracelets to magnetic mattresses. The best that can be said for such practices is that they are apparently harmless, unless the magnets get close to the patient’s computer or magnetic storage disks. Claims are made for a broad spectrum of benefits from cleansing the blood to giving the patient more energy, but clinical studies have not verified these claims, nor is there an identifiable mechanism by which such benefits might occur.

Note:
PhET Explorations: Magnet and Compass
Ever wonder how a compass worked to point you to the Arctic? Explore the interactions between a compass and bar magnet, and then add the Earth and find the surprising answer! Vary the magnet's strength, and see how

things change both inside and outside. Use the field meter to measure how the magnetic field changes.
https://archive.cnx.org/specials/5ca3e2cc-ae74-11e5-b6d3-f3c228f04b5c/magnet-and-compass/#sim-bar-magnet

Section Summary

- Crossed (perpendicular) electric and magnetic fields act as a velocity filter, giving equal and opposite forces on any charge with velocity perpendicular to the fields and of magnitude
  Equation:
  $$v = \frac{E}{B}.$$

Conceptual Questions

Exercise:

Problem:

Measurements of the weak and fluctuating magnetic fields associated with brain activity are called magnetoencephalograms (MEGs). Do the brain's magnetic fields imply coordinated or uncoordinated nerve impulses? Explain.

Exercise:

Problem:

Discuss the possibility that a Hall voltage would be generated on the moving heart of a patient during MRI imaging. Also discuss the same effect on the wires of a pacemaker. (The fact that patients with pacemakers are not given MRIs is significant.)

Exercise:

Problem:

A patient in an MRI unit turns his head quickly to one side and experiences momentary dizziness and a strange taste in his mouth. Discuss the possible causes.

Exercise:

Problem:

You are told that in a certain region there is either a uniform electric or magnetic field. What measurement or observation could you make to determine the type? (Ignore the Earth's magnetic field.)

Exercise:

Problem:

An example of magnetohydrodynamics (MHD) comes from the flow of a river (salty water). This fluid interacts with the Earth's magnetic field to produce a potential difference between the two river banks. How would you go about calculating the potential difference?

Exercise:

Problem:

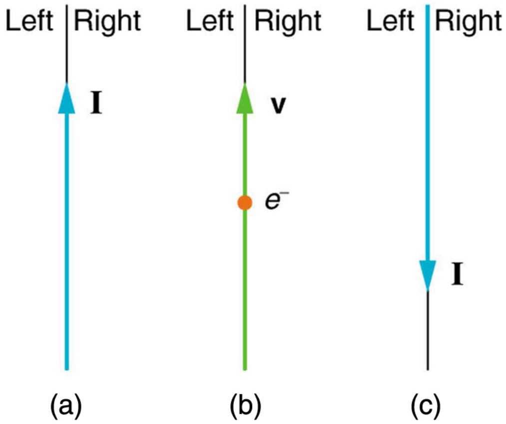
Draw gravitational field lines between 2 masses, electric field lines between a positive and a negative charge, electric field lines between 2 positive charges and magnetic field lines around a magnet. Qualitatively describe the differences between the fields and the entities responsible for the field lines.

Problems & Exercises

Exercise:

Problem:

Indicate whether the magnetic field created in each of the three situations shown in [link] is into or out of the page on the left and right of the current.



Solution:

(a) right-into page, left-out of page

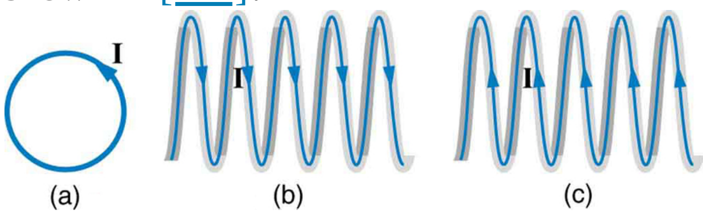
(b) right-out of page, left-into page

(c) right-out of page, left-into page

Exercise:

Problem:

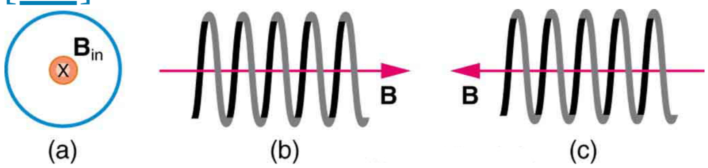
What are the directions of the fields in the center of the loop and coils shown in [link]?



Exercise:

Problem:

What are the directions of the currents in the loop and coils shown in [link]?



Solution:

(a) clockwise

(b) clockwise as seen from the left

(c) clockwise as seen from the right

Exercise:

Problem:

To see why an MRI utilizes iron to increase the magnetic field created by a coil, calculate the current needed in a 400-loop-per-meter circular coil 0.660 m in radius to create a 1.20-T field (typical of an MRI instrument) at its center with no iron present. The magnetic field of a proton is approximately like that of a circular current loop 0.650×10^{-15} m in radius carrying 1.05×10^{4} A. What is the field at the center of such a loop?

Solution:

1.01×10^{13} T

Exercise:

Problem:

Inside a motor, 30.0 A passes through a 250-turn circular loop that is 10.0 cm in radius. What is the magnetic field strength created at its center?

Exercise:

Problem:

Nonnuclear submarines use batteries for power when submerged. (a) Find the magnetic field 50.0 cm from a straight wire carrying 1200 A from the batteries to the drive mechanism of a submarine. (b) What is the field if the wires to and from the drive mechanism are side by side? (c) Discuss the effects this could have for a compass on the submarine that is not shielded.

Solution:

(a) $4.80 \times 10^{-4}\ \text{T}$

(b) Zero

(c) If the wires are not paired, the field is about 10 times stronger than Earth's magnetic field and so could severely disrupt the use of a compass.

Exercise:

Problem:

How strong is the magnetic field inside a solenoid with 10,000 turns per meter that carries 20.0 A?

Exercise:

Problem:

What current is needed in the solenoid described in [link] to produce a magnetic field 10^4 times the Earth's magnetic field of $5.00 \times 10^{-5}\ \text{T}$?

Solution:

39.8 A

Exercise:

Problem:

How far from the starter cable of a car, carrying 150 A, must you be to experience a field less than the Earth's $(5.00 \times 10^{-5}\ \text{T})$? Assume a long straight wire carries the current. (In practice, the body of your car shields the dashboard compass.)

Exercise:

Problem:

Measurements affect the system being measured, such as the current loop in [link]. (a) Estimate the field the loop creates by calculating the field at the center of a circular loop 20.0 cm in diameter carrying 5.00 A. (b) What is the smallest field strength this loop can be used to measure, if its field must alter the measured field by less than 0.0100%?

Solution:

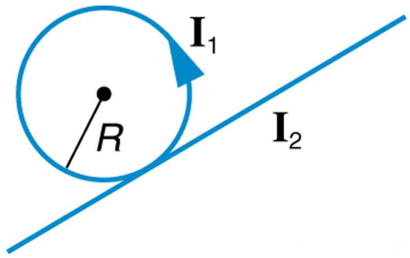
(a) $3.14 \times 10^{-5}\ \text{T}$

(b) 0.314 T

Exercise:

Problem:

[link] shows a long straight wire just touching a loop carrying a current I_1. Both lie in the same plane. (a) What direction must the current I_2 in the straight wire have to create a field at the center of the loop in the direction opposite to that created by the loop? (b) What is the ratio of I_1/I_2 that gives zero field strength at the center of the loop? (c) What is the direction of the field directly above the loop under this circumstance?



Exercise:

Problem:

Find the magnitude and direction of the magnetic field at the point equidistant from the wires in [link](a), using the rules of vector addition to sum the contributions from each wire.

Solution:

7.55×10^{-5} T, 23.4º

Exercise:

Problem:

Find the magnitude and direction of the magnetic field at the point equidistant from the wires in [link](b), using the rules of vector addition to sum the contributions from each wire.

Exercise:

Problem:

What current is needed in the top wire in [link](a) to produce a field of zero at the point equidistant from the wires, if the currents in the bottom two wires are both 10.0 A into the page?

Solution:

10.0 A

Exercise:

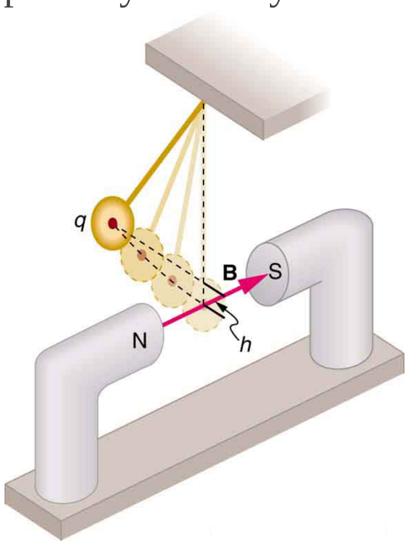
Problem:

Calculate the size of the magnetic field 20 m below a high voltage power line. The line carries 450 MW at a voltage of 300,000 V.

Exercise:

Problem: Integrated Concepts

(a) A pendulum is set up so that its bob (a thin copper disk) swings between the poles of a permanent magnet as shown in [link]. What is the magnitude and direction of the magnetic force on the bob at the lowest point in its path, if it has a positive $0.250\ \mu C$ charge and is released from a height of 30.0 cm above its lowest point? The magnetic field strength is 1.50 T. (b) What is the acceleration of the bob at the bottom of its swing if its mass is 30.0 grams and it is hung from a flexible string? Be certain to include a free-body diagram as part of your analysis.



Solution:

(a) 9.09×10^{-7} N upward

(b) $3.03 \times 10^{-5}\ \mathrm{m/s^2}$

Exercise:

Problem: **Integrated Concepts**

(a) What voltage will accelerate electrons to a speed of $6.00 \times 10^{-7}\ \text{m/s}$? (b) Find the radius of curvature of the path of a *proton* accelerated through this potential in a 0.500-T field and compare this with the radius of curvature of an electron accelerated through the same potential.

Exercise:

Problem: **Integrated Concepts**

Find the radius of curvature of the path of a 25.0-MeV proton moving perpendicularly to the 1.20-T field of a cyclotron.

Solution:

60.2 cm

Exercise:

Problem: **Integrated Concepts**

To construct a nonmechanical water meter, a 0.500-T magnetic field is placed across the supply water pipe to a home and the Hall voltage is recorded. (a) Find the flow rate in liters per second through a 3.00-cm-diameter pipe if the Hall voltage is 60.0 mV. (b) What would the Hall voltage be for the same flow rate through a 10.0-cm-diameter pipe with the same field applied?

Exercise:

Problem: **Integrated Concepts**

(a) Using the values given for an MHD drive in [link], and assuming the force is uniformly applied to the fluid, calculate the pressure created in N/m^2. (b) Is this a significant fraction of an atmosphere?

Solution:

(a) $1.02 \times 10^3\ \mathrm{N/m}^2$

(b) Not a significant fraction of an atmosphere

Exercise:

Problem: Integrated Concepts

(a) Calculate the maximum torque on a 50-turn, 1.50 cm radius circular current loop carrying $50\ \mu\mathrm{A}$ in a 0.500-T field. (b) If this coil is to be used in a galvanometer that reads $50\ \mu\mathrm{A}$ full scale, what force constant spring must be used, if it is attached 1.00 cm from the axis of rotation and is stretched by the 60^o arc moved?

Exercise:

Problem: Integrated Concepts

A current balance used to define the ampere is designed so that the current through it is constant, as is the distance between wires. Even so, if the wires change length with temperature, the force between them will change. What percent change in force per degree will occur if the wires are copper?

Solution:

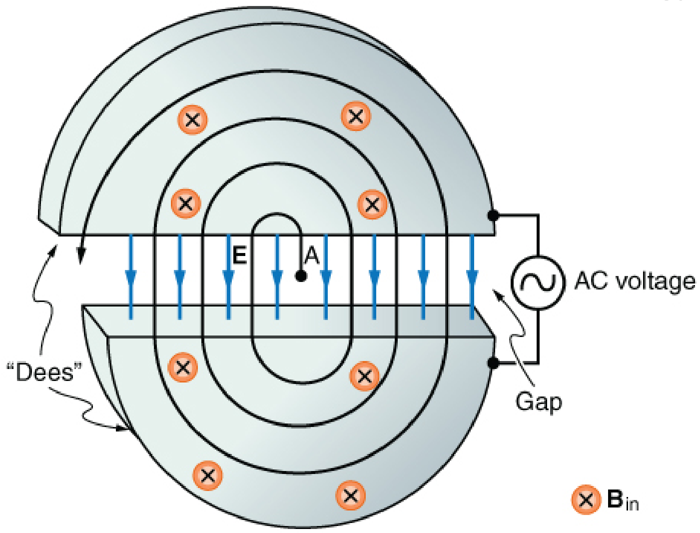
$17.0 \times 10^{-4}\%/^\mathrm{o}\mathrm{C}$

Exercise:

Problem:Integrated Concepts

(a) Show that the period of the circular orbit of a charged particle moving perpendicularly to a uniform magnetic field is $T = 2\pi m/(\mathrm{qB})$. (b) What is the frequency f? (c) What is the angular

velocity ω? Note that these results are independent of the velocity and radius of the orbit and, hence, of the energy of the particle. ([link].)



Cyclotrons accelerate charged particles orbiting in a magnetic field by placing an AC voltage on the metal Dees, between which the particles move, so that energy is added twice each orbit. The frequency is constant, since it is independent of the particle energy—the radius of the orbit simply increases with energy until the particles approach the edge and are extracted for various experiments and applications.

Exercise:

Problem: Integrated Concepts

A cyclotron accelerates charged particles as shown in [link]. Using the results of the previous problem, calculate the frequency of the accelerating voltage needed for a proton in a 1.20-T field.

Solution:

18.3 MHz

Exercise:

Problem: Integrated Concepts

(a) A 0.140-kg baseball, pitched at 40.0 m/s horizontally and perpendicular to the Earth's horizontal $5.00 \times 10^{-5}\ \mathrm{T}$ field, has a 100-nC charge on it. What distance is it deflected from its path by the magnetic force, after traveling 30.0 m horizontally? (b) Would you suggest this as a secret technique for a pitcher to throw curve balls?

Exercise:

Problem: Integrated Concepts

(a) What is the direction of the force on a wire carrying a current due east in a location where the Earth's field is due north? Both are parallel to the ground. (b) Calculate the force per meter if the wire carries 20.0 A and the field strength is $3.00 \times 10^{-5}\ \mathrm{T}$. (c) What diameter copper wire would have its weight supported by this force? (d) Calculate the resistance per meter and the voltage per meter needed.

Solution:

(a) Straight up

(b) $6.00 \times 10^{-4}\ \mathrm{N/m}$

(c) 94.1 μm

(d)2.47 Ω/m, 49.4 V/m

Exercise:

Problem: Integrated Concepts

One long straight wire is to be held directly above another by repulsion between their currents. The lower wire carries 100 A and the wire 7.50 cm above it is 10-gauge (2.588 mm diameter) copper wire. (a) What current must flow in the upper wire, neglecting the Earth's field? (b) What is the smallest current if the Earth's $3.00 \times 10^{-5}\ \mathrm{T}$ field is parallel to the ground and is not neglected? (c) Is the supported wire in a stable or unstable equilibrium if displaced vertically? If displaced horizontally?

Exercise:

Problem: Unreasonable Results

(a) Find the charge on a baseball, thrown at 35.0 m/s perpendicular to the Earth's $5.00 \times 10^{-5}\ \mathrm{T}$ field, that experiences a 1.00-N magnetic force. (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Solution:

(a) 571 C

(b) Impossible to have such a large separated charge on such a small object.

(c) The 1.00-N force is much too great to be realistic in the Earth's field.

Exercise:

Problem: Unreasonable Results

A charged particle having mass 6.64×10^{-27} kg (that of a helium atom) moving at $8.70 \times 10^{5}\ \mathrm{m/s}$ perpendicular to a 1.50-T magnetic field travels in a circular path of radius 16.0 mm. (a) What is the charge of the particle? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Exercise:

Problem: **Unreasonable Results**

An inventor wants to generate 120-V power by moving a 1.00-m-long wire perpendicular to the Earth's $5.00 \times 10^{-5}\ \mathrm{T}$ field. (a) Find the speed with which the wire must move. (b) What is unreasonable about this result? (c) Which assumption is responsible?

Solution:

(a) $2.40 \times 10^{6}\ \mathrm{m/s}$

(b) The speed is too high to be practical $\leq$ 1% speed of light

(c) The assumption that you could reasonably generate such a voltage with a single wire in the Earth's field is unreasonable

Exercise:

Problem: **Unreasonable Results**

Frustrated by the small Hall voltage obtained in blood flow measurements, a medical physicist decides to increase the applied magnetic field strength to get a 0.500-V output for blood moving at 30.0 cm/s in a 1.50-cm-diameter vessel. (a) What magnetic field strength is needed? (b) What is unreasonable about this result? (c) Which premise is responsible?

Exercise:

Problem: **Unreasonable Results**

A surveyor 100 m from a long straight 200-kV DC power line suspects that its magnetic field may equal that of the Earth and affect compass readings. (a) Calculate the current in the wire needed to create a $5.00 \times 10^{-5}\ \mathrm{T}$ field at this distance. (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Solution:

(a) 25.0 kA

(b) This current is unreasonably high. It implies a total power delivery in the line of 50.0x10^9 W, which is much too high for standard transmission lines.

(c)100 meters is a long distance to obtain the required field strength. Also coaxial cables are used for transmission lines so that there is virtually no field for DC power lines, because of cancellation from opposing currents. The surveyor's concerns are not a problem for his magnetic field measurements.

Exercise:

Problem:Construct Your Own Problem

Consider a mass separator that applies a magnetic field perpendicular to the velocity of ions and separates the ions based on the radius of curvature of their paths in the field. Construct a problem in which you calculate the magnetic field strength needed to separate two ions that differ in mass, but not charge, and have the same initial velocity. Among the things to consider are the types of ions, the velocities they can be given before entering the magnetic field, and a reasonable value for the radius of curvature of the paths they follow. In addition, calculate the separation distance between the ions at the point where they are detected.

Exercise:

Problem: Construct Your Own Problem

Consider using the torque on a current-carrying coil in a magnetic field to detect relatively small magnetic fields (less than the field of the Earth, for example). Construct a problem in which you calculate the maximum torque on a current-carrying loop in a magnetic field. Among the things to be considered are the size of the coil, the number

of loops it has, the current you pass through the coil, and the size of the field you wish to detect. Discuss whether the torque produced is large enough to be effectively measured. Your instructor may also wish for you to consider the effects, if any, of the field produced by the coil on the surroundings that could affect detection of the small field.

Glossary

magnetic resonance imaging (MRI)
: a medical imaging technique that uses magnetic fields create detailed images of internal tissues and organs

nuclear magnetic resonance (NMR)
: a phenomenon in which an externally applied magnetic field interacts with the nuclei of certain atoms

magnetocardiogram (MCG)
: a recording of the heart’s magnetic field as it beats

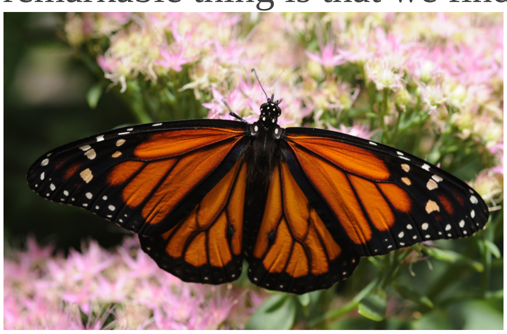
magnetoencephalogram (MEG)
: a measurement of the brain’s magnetic field

Introduction to Electromagnetic Induction, AC Circuits and Electrical Technologies

class="introduction"

These wind turbines in the Thames Estuary in the UK are an example of induction at work. Wind pushes the blades of the turbine, spinning a shaft attached to magnets. The magnets spin around a conductive coil, inducing an electric current in the coil, and eventually feeding the electrical grid. (credit: modification of work by Petr Kratochvil)

Nature’s displays of symmetry are beautiful and alluring. A butterfly’s wings exhibit an appealing symmetry in a complex system. (See [link].) The laws of physics display symmetries at the most basic level—these symmetries are a source of wonder and imply deeper meaning. Since we place a high value on symmetry, we look for it when we explore nature. The remarkable thing is that we find it.



Physics, like this butterfly, has inherent symmetries. (credit: Thomas Bresson)

The hint of symmetry between electricity and magnetism found in the preceding chapter will be elaborated upon in this chapter. Specifically, we know that a current creates a magnetic field. If nature is symmetric here, then perhaps a magnetic field can create a current. The Hall effect is a voltage caused by a magnetic force. That voltage could drive a current. Historically, it was very shortly after Oersted discovered currents cause magnetic fields that other scientists asked the following question: Can magnetic fields cause currents? The answer was soon found by experiment to be yes. In 1831, some 12 years after Oersted's discovery, the English scientist Michael Faraday (1791–1862) and the American scientist Joseph Henry (1797–1878) independently demonstrated that magnetic fields can produce currents. The basic process of generating emfs (electromotive force) and, hence, currents with magnetic fields is known as **induction**; this process is also called magnetic induction to distinguish it from charging by induction, which utilizes the Coulomb force.

Today, currents induced by magnetic fields are essential to our technological society. The ubiquitous generator—found in automobiles, on bicycles, in nuclear power plants, and so on—uses magnetism to generate current. Other devices that use magnetism to induce currents include pickup coils in electric guitars, transformers of every size, certain microphones, airport security gates, and damping mechanisms on sensitive chemical balances. Not so familiar perhaps, but important nevertheless, is that the behavior of AC circuits depends strongly on the effect of magnetic fields on currents.

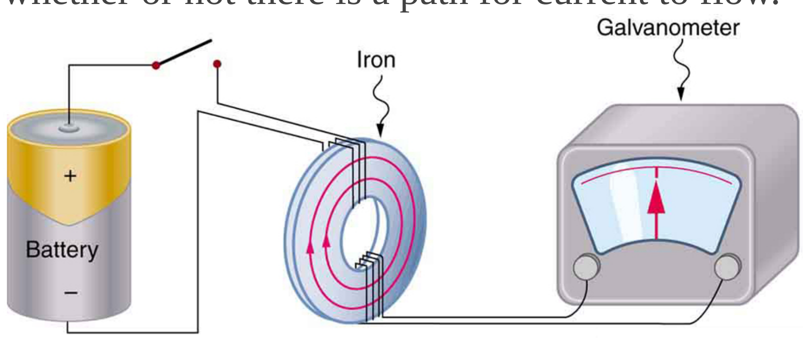
Glossary

induction
: (magnetic induction) the creation of emfs and hence currents by magnetic fields

Induced Emf and Magnetic Flux

- Calculate the flux of a uniform magnetic field through a loop of arbitrary orientation.
- Describe methods to produce an electromotive force (emf) with a magnetic field or magnet and a loop of wire.

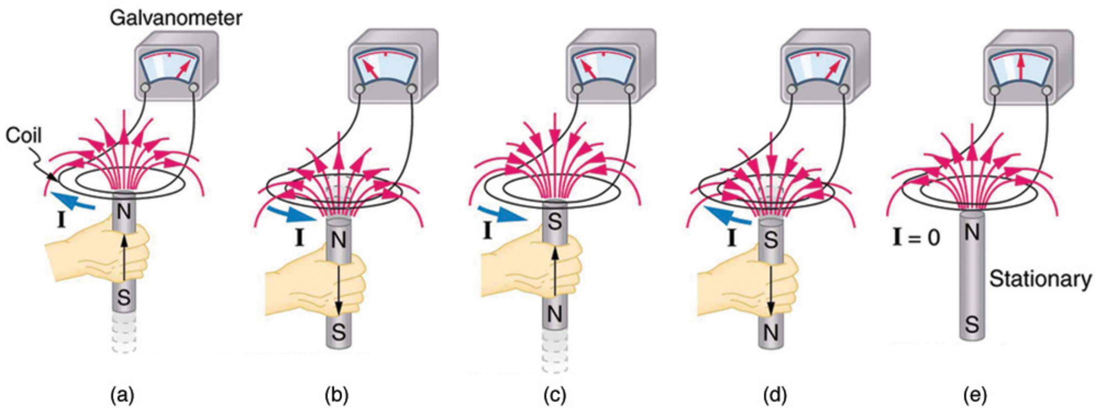
The apparatus used by Faraday to demonstrate that magnetic fields can create currents is illustrated in [link]. When the switch is closed, a magnetic field is produced in the coil on the top part of the iron ring and transmitted to the coil on the bottom part of the ring. The galvanometer is used to detect any current induced in the coil on the bottom. It was found that each time the switch is closed, the galvanometer detects a current in one direction in the coil on the bottom. (You can also observe this in a physics lab.) Each time the switch is opened, the galvanometer detects a current in the opposite direction. Interestingly, if the switch remains closed or open for any length of time, there is no current through the galvanometer. *Closing and opening the switch* induces the current. It is the *change* in magnetic field that creates the current. More basic than the current that flows is the emfthat causes it. The current is a result of an *emf induced by a changing magnetic field*, whether or not there is a path for current to flow.



Faraday’s apparatus for demonstrating that a magnetic field can produce a current. A change in the field produced by the top coil induces an emf and, hence, a current in the bottom coil. When the switch is opened and closed, the galvanometer registers currents in opposite directions. No current flows

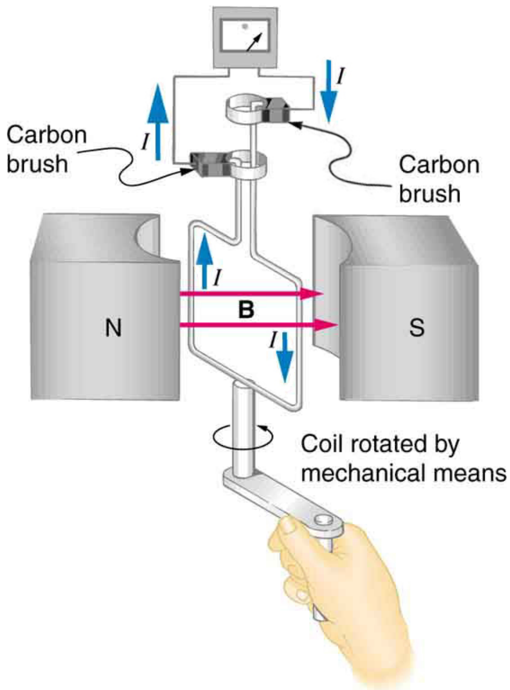
through the galvanometer when the switch remains closed or open.

An experiment easily performed and often done in physics labs is illustrated in [link]. An emf is induced in the coil when a bar magnet is pushed in and out of it. Emfs of opposite signs are produced by motion in opposite directions, and the emfs are also reversed by reversing poles. The same results are produced if the coil is moved rather than the magnet—it is the relative motion that is important. The faster the motion, the greater the emf, and there is no emf when the magnet is stationary relative to the coil.



Movement of a magnet relative to a coil produces emfs as shown. The same emfs are produced if the coil is moved relative to the magnet. The greater the speed, the greater the magnitude of the emf, and the emf is zero when there is no motion.

The method of inducing an emf used in most electric generators is shown in [link]. A coil is rotated in a magnetic field, producing an alternating current emf, which depends on rotation rate and other factors that will be explored in later sections. Note that the generator is remarkably similar in construction to a motor (another symmetry).



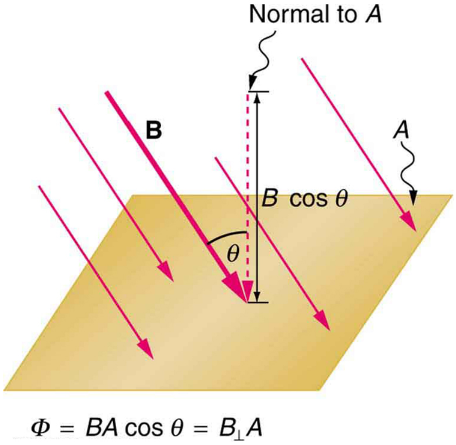
Rotation of a coil in a magnetic field produces an emf. This is the basic construction of a generator, where work done to turn the coil is converted to electric energy. Note the generator is very similar in construction to a motor.

So we see that changing the magnitude or direction of a magnetic field produces an emf. Experiments revealed that there is a crucial quantity called the **magnetic flux**, Φ , given by

Equation:

$$\Phi = \mathrm{BA}\cos\theta,$$

where B is the magnetic field strength over an area A, at an angle θ with the perpendicular to the area as shown in [link]. **Any change in magnetic flux Φ induces an emf.** This process is defined to be **electromagnetic induction**. Units of magnetic flux Φ are $\mathrm{T \cdot m^2}$. As seen in [link], $B \cos \theta = B_{\perp}$, which is the component of B perpendicular to the area A. Thus magnetic flux is $\Phi = B_{\perp} A$, the product of the area and the component of the magnetic field perpendicular to it.



Magnetic flux Φ is related to the magnetic field and the area over which it exists. The flux $\Phi = \mathrm{BA} \cos \theta$ is related to induction; any change in Φ induces an emf.

All induction, including the examples given so far, arises from some change in magnetic flux Φ. For example, Faraday changed B and hence Φ when opening and closing the switch in his apparatus (shown in [link]). This is also true for the bar magnet and coil shown in [link]. When rotating the coil of a generator, the angle θ and, hence, Φ is changed. Just how great an emf and what direction it takes depend on the change in Φ and how rapidly the change is made, as examined in the next section.

Section Summary

- The crucial quantity in induction is magnetic flux Φ, defined to be $\Phi = \text{BA}\cos\theta$, where B is the magnetic field strength over an area A at an angle θ with the perpendicular to the area.
- Units of magnetic flux Φ are $\text{T}\cdot\text{m}^2$.
- Any change in magnetic flux Φ induces an emf—the process is defined to be electromagnetic induction.

Conceptual Questions

Exercise:

Problem:

How do the multiple-loop coils and iron ring in the version of Faraday's apparatus shown in [link] enhance the observation of induced emf?

Exercise:

Problem:

When a magnet is thrust into a coil as in [link](a), what is the direction of the force exerted by the coil on the magnet? Draw a diagram showing the direction of the current induced in the coil and the magnetic field it produces, to justify your response. How does the magnitude of the force depend on the resistance of the galvanometer?

Exercise:

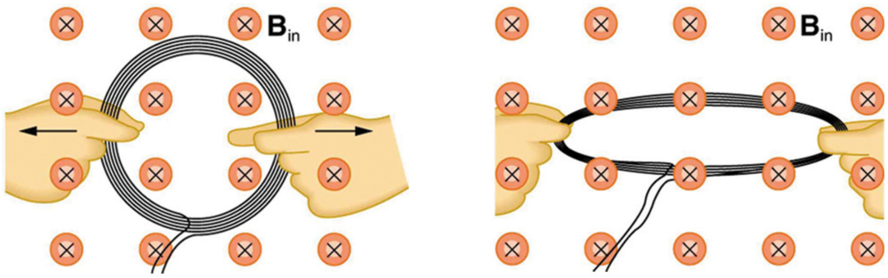
Problem:

Explain how magnetic flux can be zero when the magnetic field is not zero.

Exercise:

Problem:

Is an emf induced in the coil in [link] when it is stretched? If so, state why and give the direction of the induced current.



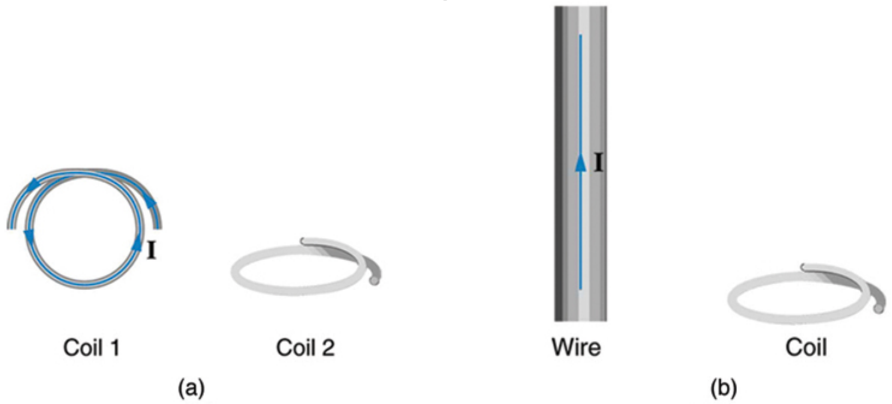
A circular coil of wire is stretched in a magnetic field.

Problems & Exercises

Exercise:

Problem:

What is the value of the magnetic flux at coil 2 in [link] due to coil 1?



(a) The planes of the two coils are perpendicular. (b) The wire is perpendicular to the plane of the coil.

Solution:

Zero

Exercise:

Problem:

What is the value of the magnetic flux through the coil in [link](b) due to the wire?

Glossary

magnetic flux
: the amount of magnetic field going through a particular area, calculated with $\Phi = \mathrm{BA}\cos\theta$ where B is the magnetic field strength over an area A at an angle θ with the perpendicular to the area

electromagnetic induction
: the process of inducing an emf (voltage) with a change in magnetic flux

Faraday’s Law of Induction: Lenz’s Law

- Calculate emf, current, and magnetic fields using Faraday’s Law.
- Explain the physical results of Lenz’s Law

Faraday’s and Lenz’s Law

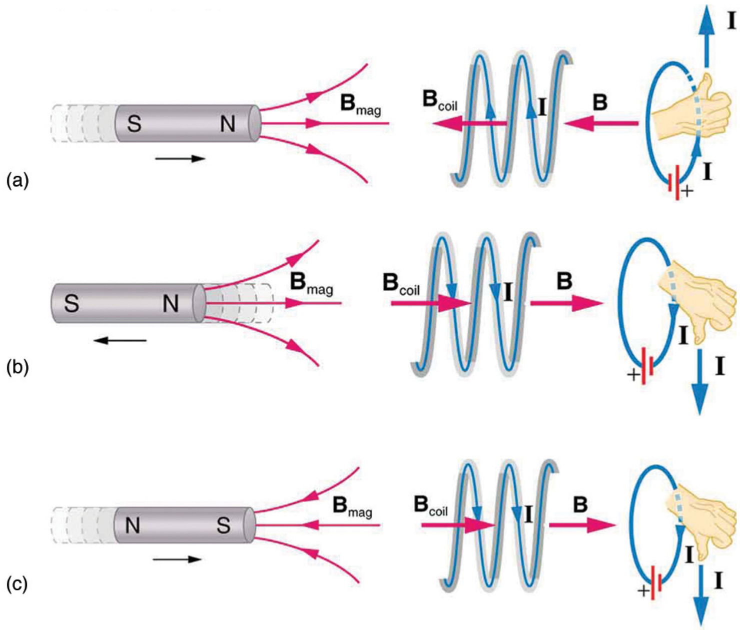
Faraday’s experiments showed that the emf induced by a change in magnetic flux depends on only a few factors. First, emf is directly proportional to the change in flux $\Delta\Phi$. Second, emf is greatest when the change in time Δt is smallest—that is, emf is inversely proportional to Δt. Finally, if a coil has N turns, an emf will be produced that is N times greater than for a single coil, so that emf is directly proportional to N. The equation for the emf induced by a change in magnetic flux is

Equation:

$$\text{emf} = -N\frac{\Delta\Phi}{\Delta t}.$$

This relationship is known as **Faraday’s law of induction**. The units for emf are volts, as is usual.

The minus sign in Faraday’s law of induction is very important. The minus means that *the emf creates a current I and magnetic field B that oppose the change in flux $\Delta\Phi$—this is known as Lenz’s law*. The direction (given by the minus sign) of the emfis so important that it is called **Lenz’s law** after the Russian Heinrich Lenz (1804–1865), who, like Faraday and Henry,independently investigated aspects of induction. Faraday was aware of the direction, but Lenz stated it so clearly that he is credited for its discovery. (See [link].)



(a) When this bar magnet is thrust into the coil, the strength of the magnetic field increases in the coil. The current induced in the coil creates another field, in the opposite direction of the bar magnet's to oppose the increase. This is one aspect of *Lenz's law—induction opposes any change in flux*. (b) and (c) are two other situations. Verify for yourself that the direction of the induced B_{coil} shown indeed opposes the change in flux and that the current direction shown is consistent with RHR-2.

Note:

Problem-Solving Strategy for Lenz's Law

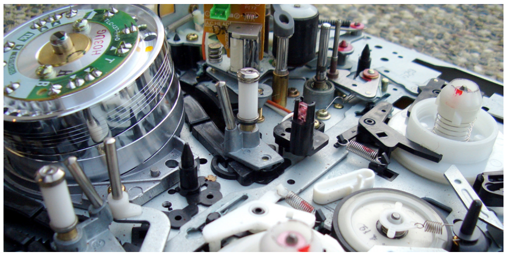
To use Lenz's law to determine the directions of the induced magnetic fields, currents, and emfs:

1. Make a sketch of the situation for use in visualizing and recording directions.
2. Determine the direction of the magnetic field B.
3. Determine whether the flux is increasing or decreasing.
4. Now determine the direction of the induced magnetic field B. It opposes the *change* in flux by adding or subtracting from the original field.
5. Use RHR-2 to determine the direction of the induced current I that is responsible for the induced magnetic field B.
6. The direction (or polarity) of the induced emf will now drive a current in this direction and can be represented as current emerging from the positive terminal of the emf and returning to its negative terminal.

For practice, apply these steps to the situations shown in [link] and to others that are part of the following text material.

Applications of Electromagnetic Induction

There are many applications of Faraday's Law of induction, as we will explore in this chapter and others. At this juncture, let us mention several that have to do with data storage and magnetic fields. A very important application has to do with audio and video *recording tapes*. A plastic tape, coated with iron oxide, moves past a recording head. This recording head is basically a round iron ring about which is wrapped a coil of wire—an electromagnet ([link]). A signal in the form of a varying input current from a microphone or camera goes to the recording head. These signals (which are a function of the signal amplitude and frequency) produce varying magnetic fields at the recording head. As the tape moves past the recording head, the magnetic field orientations of the iron oxide molecules on the tape are changed thus recording the signal. In the playback mode, the magnetized tape is run past another head, similar in structure to the recording head. The different magnetic field orientations of the iron oxide molecules on the tape induces an emf in the coil of wire in the playback head. This signal then is sent to a loudspeaker or video player.



Recording and playback heads used with audio and video magnetic tapes. (credit: Steve Jurvetson)

Similar principles apply to computer hard drives, except at a much faster rate. Here recordings are on a coated, spinning disk. Read heads historically were made to work on the principle of induction. However, the input information is carried in digital rather than analog form – a series of 0’s or 1’s are written upon the spinning hard drive. Today, most hard drive readout devices do not work on the principle of induction, but use a technique known as *giant magnetoresistance*. (The discovery that weak changes in a magnetic field in a thin film of iron and chromium could bring about much larger changes in electrical resistance was one of the first large successes of nanotechnology.) Another application of induction is found on the magnetic stripe on the back of your personal credit card as used at the grocery store or the ATM machine. This works on the same principle as the audio or video tape mentioned in the last paragraph in which a head reads personal information from your card.

Another application of electromagnetic induction is when electrical signals need to be transmitted across a barrier. Consider the *cochlear implant* shown below. Sound is picked up by a microphone on the outside of the skull and is used to set up a varying magnetic field. A current is induced in a receiver secured in the bone beneath the skin and transmitted to electrodes in the inner ear. Electromagnetic induction can be used in other instances where electric signals need to be conveyed across various media.

Electromagnetic induction used in transmitting electric currents across mediums. The device on the baby’s head induces an electrical current in a receiver secured in the bone beneath the skin. (credit: Bjorn Knetsch)

Another contemporary area of research in which electromagnetic induction is being successfully implemented (and with substantial potential) is transcranial magnetic simulation. A host of disorders, including depression and hallucinations can be traced to irregular localized electrical activity in the brain. In *transcranial magnetic stimulation*, a rapidly varying and very localized magnetic field is placed close to certain sites identified in the brain. Weak electric currents are induced in the identified sites and can result in recovery of electrical functioning in the brain tissue.

Sleep apnea (“the cessation of breath”) affects both adults and infants (especially premature babies and it may be a cause of sudden infant deaths [SID]). In such individuals, breath can stop repeatedly during their sleep. A cessation of more than 20 seconds can be very dangerous. Stroke, heart failure, and tiredness are just some of the possible consequences for a

person having sleep apnea. The concern in infants is the stopping of breath for these longer times. One type of monitor to alert parents when a child is not breathing uses electromagnetic induction. A wire wrapped around the infant's chest has an alternating current running through it. The expansion and contraction of the infant's chest as the infant breathes changes the area through the coil. A pickup coil located nearby has an alternating current induced in it due to the changing magnetic field of the initial wire. If the child stops breathing, there will be a change in the induced current, and so a parent can be alerted.

Note:

Making Connections: Conservation of Energy

Lenz's law is a manifestation of the conservation of energy. The induced emf produces a current that opposes the change in flux, because a change in flux means a change in energy. Energy can enter or leave, but not instantaneously. Lenz's law is a consequence. As the change begins, the law says induction opposes and, thus, slows the change. In fact, if the induced emf were in the same direction as the change in flux, there would be a positive feedback that would give us free energy from no apparent source—conservation of energy would be violated.

Example:

Calculating Emf: How Great Is the Induced Emf?

Calculate the magnitude of the induced emf when the magnet in [link](a) is thrust into the coil, given the following information: the single loop coil has a radius of 6.00 cm and the average value of $B \cos \theta$ (this is given, since the bar magnet's field is complex) increases from 0.0500 T to 0.250 T in 0.100 s.

Strategy

To find the *magnitude* of emf, we use Faraday's law of induction as stated by $\text{emf} = -N\frac{\Delta \Phi}{\Delta t}$, but without the minus sign that indicates direction:

Equation:

$$\text{emf} = N\frac{\Delta\Phi}{\Delta t}.$$

Solution
We are given that $N = 1$ and $\Delta t = 0.100$ s, but we must determine the change in flux $\Delta\Phi$ before we can find emf. Since the area of the loop is fixed, we see that
Equation:

$$\Delta\Phi = \Delta(BA\ \cos\theta) = A\Delta(B\cos\theta\ ).$$

Now $\Delta(B\cos\theta) = 0.200$ T, since it was given that $B\cos\theta$ changes from 0.0500 to 0.250 T. The area of the loop is $A = \pi r^2 = (3.14...)(0.060\ \text{m})^2 = 1.13 \times 10^{-2}\ \text{m}^2$. Thus,
Equation:

$$\Delta\Phi = (1.13 \times 10^{-2}\ \text{m}^2)(0.200\ \text{T}\ ).$$

Entering the determined values into the expression for emf gives
Equation:

$$\text{Emf} = N\frac{\Delta\Phi}{\Delta t} = \frac{(1.13 \times 10^{-2}\ \text{m}^2)(0.200\ \text{T})}{0.100\ \text{s}} = 22.6\ \text{mV}.$$

Discussion
While this is an easily measured voltage, it is certainly not large enough for most practical applications. More loops in the coil, a stronger magnet, and faster movement make induction the practical source of voltages that it is.

Note:
PhET Explorations: Faraday's Electromagnetic Lab
Play with a bar magnet and coils to learn about Faraday's law. Move a bar magnet near one or two coils to make a light bulb glow. View the magnetic field lines. A meter shows the direction and magnitude of the current. View the magnetic field lines or use a meter to show the direction and magnitude of the current. You can also play with electromagnets, generators and

transformers!

https://archive.cnx.org/specials/70b14c10-ae73-11e5-8eb2-b7fbe0c5c7a4/faraday/#sim-bar-magnet

Section Summary

- Faraday's law of induction states that the emfinduced by a change in magnetic flux is
  Equation:
  $$\text{emf} = -N\frac{\Delta \Phi}{\Delta t}$$
  when flux changes by $\Delta \Phi$ in a time Δt.
- If emf is induced in a coil, N is its number of turns.
- The minus sign means that the emf creates a current I and magnetic field B that *oppose the change in flux* $\Delta \Phi$ —this opposition is known as Lenz's law.

Conceptual Questions

Exercise:

Problem:

A person who works with large magnets sometimes places her head inside a strong field. She reports feeling dizzy as she quickly turns her head. How might this be associated with induction?

Exercise:

Problem:

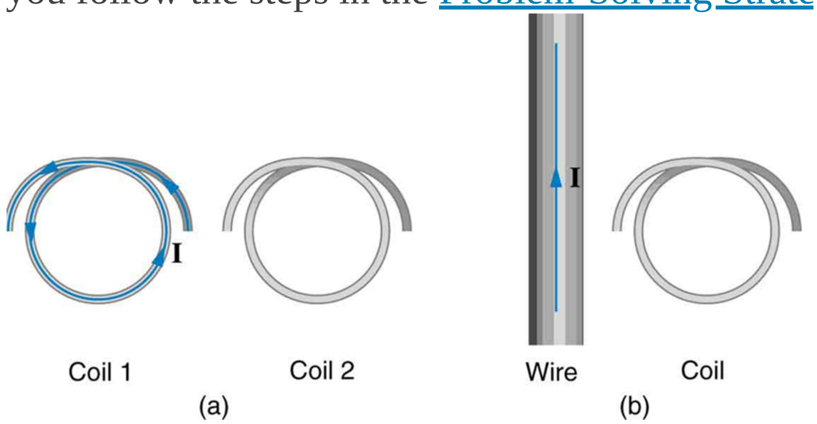
A particle accelerator sends high-velocity charged particles down an evacuated pipe. Explain how a coil of wire wrapped around the pipe could detect the passage of individual particles. Sketch a graph of the voltage output of the coil as a single particle passes through it.

Problems & Exercises

Exercise:

Problem:

Referring to [link](a), what is the direction of the current induced in coil 2: (a) If the current in coil 1 increases? (b) If the current in coil 1 decreases? (c) If the current in coil 1 is constant? Explicitly show how you follow the steps in the Problem-Solving Strategy for Lenz's Law.



(a) The coils lie in the same plane. (b) The wire is in the plane of the coil

---

Solution:

(a) CCW

(b) CW

(c) No current induced

Exercise:

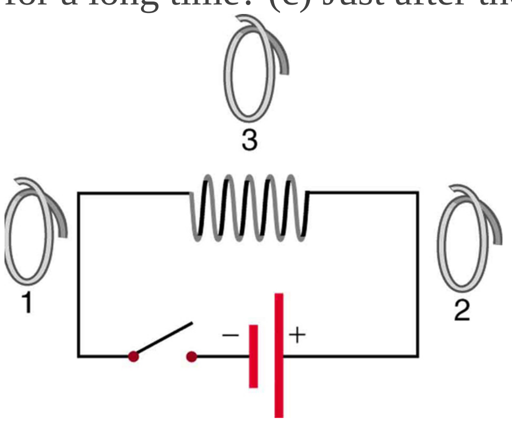
Problem:

Referring to [link](b), what is the direction of the current induced in the coil: (a) If the current in the wire increases? (b) If the current in the wire decreases? (c) If the current in the wire suddenly changes direction? Explicitly show how you follow the steps in the Problem-Solving Strategy for Lenz’s Law.

Exercise:

Problem:

Referring to [link], what are the directions of the currents in coils 1, 2, and 3 (assume that the coils are lying in the plane of the circuit): (a) When the switch is first closed? (b) When the switch has been closed for a long time? (c) Just after the switch is opened?



Solution:

(a) 1 CCW, 2 CCW, 3 CW

(b) 1, 2, and 3 no current induced

(c) 1 CW, 2 CW, 3 CCW

Exercise:

Problem: Repeat the previous problem with the battery reversed.

Exercise:

Problem:

Verify that the units of $\Delta\Phi / \Delta t$ are volts. That is, show that $1\ \mathrm{T} \cdot \mathrm{m}^2/\mathrm{s} = 1\ \mathrm{V}$.

Exercise:

Problem:

Suppose a 50-turn coil lies in the plane of the page in a uniform magnetic field that is directed into the page. The coil originally has an area of $0.250\ \mathrm{m}^2$. It is stretched to have no area in 0.100 s. What is the direction and magnitude of the induced emf if the uniform magnetic field has a strength of 1.50 T?

Exercise:

Problem:

(a) An MRI technician moves his hand from a region of very low magnetic field strength into an MRI scanner's 2.00 T field with his fingers pointing in the direction of the field. Find the average emf induced in his wedding ring, given its diameter is 2.20 cm and assuming it takes 0.250 s to move it into the field. (b) Discuss whether this current would significantly change the temperature of the ring.

Solution:

(a) 3.04 mV
(b) As a lower limit on the ring, estimate R = 1.00 mΩ. The heat transferred will be 2.31 mJ. This is not a significant amount of heat.

Exercise:

Problem: Integrated Concepts

Referring to the situation in the previous problem: (a) What current is induced in the ring if its resistance is 0.0100 Ω? (b) What average power is dissipated? (c) What magnetic field is induced at the center of the ring? (d) What is the direction of the induced magnetic field relative to the MRI's field?

Exercise:

Problem:

An emf is induced by rotating a 1000-turn, 20.0 cm diameter coil in the Earth's 5.00×10^{-5} T magnetic field. What average emf is induced, given the plane of the coil is originally perpendicular to the Earth's field and is rotated to be parallel to the field in 10.0 ms?

Solution:

0.157 V

Exercise:

Problem:

A 0.250 m radius, 500-turn coil is rotated one-fourth of a revolution in 4.17 ms, originally having its plane perpendicular to a uniform magnetic field. (This is 60 rev/s.) Find the magnetic field strength needed to induce an average emf of 10,000 V.

Exercise:

Problem: Integrated Concepts

Approximately how does the emf induced in the loop in [link](b) depend on the distance of the center of the loop from the wire?

Solution:

proportional to $\frac{1}{r}$

Exercise:

Problem: Integrated Concepts

(a) A lightning bolt produces a rapidly varying magnetic field. If the bolt strikes the earth vertically and acts like a current in a long straight wire, it will induce a voltage in a loop aligned like that in [link](b). What voltage is induced in a 1.00 m diameter loop 50.0 m from a 2.00×10^{6} A lightning strike, if the current falls to zero in 25.0 μs? (b) Discuss circumstances under which such a voltage would produce noticeable consequences.

Glossary

Faraday’s law of induction
: the means of calculating the emf in a coil due to changing magnetic flux, given by $\text{emf} = -N\frac{\Delta\Phi}{\Delta t}$

Lenz’s law
: the minus sign in Faraday’s law, signifying that the emf induced in a coil opposes the change in magnetic flux

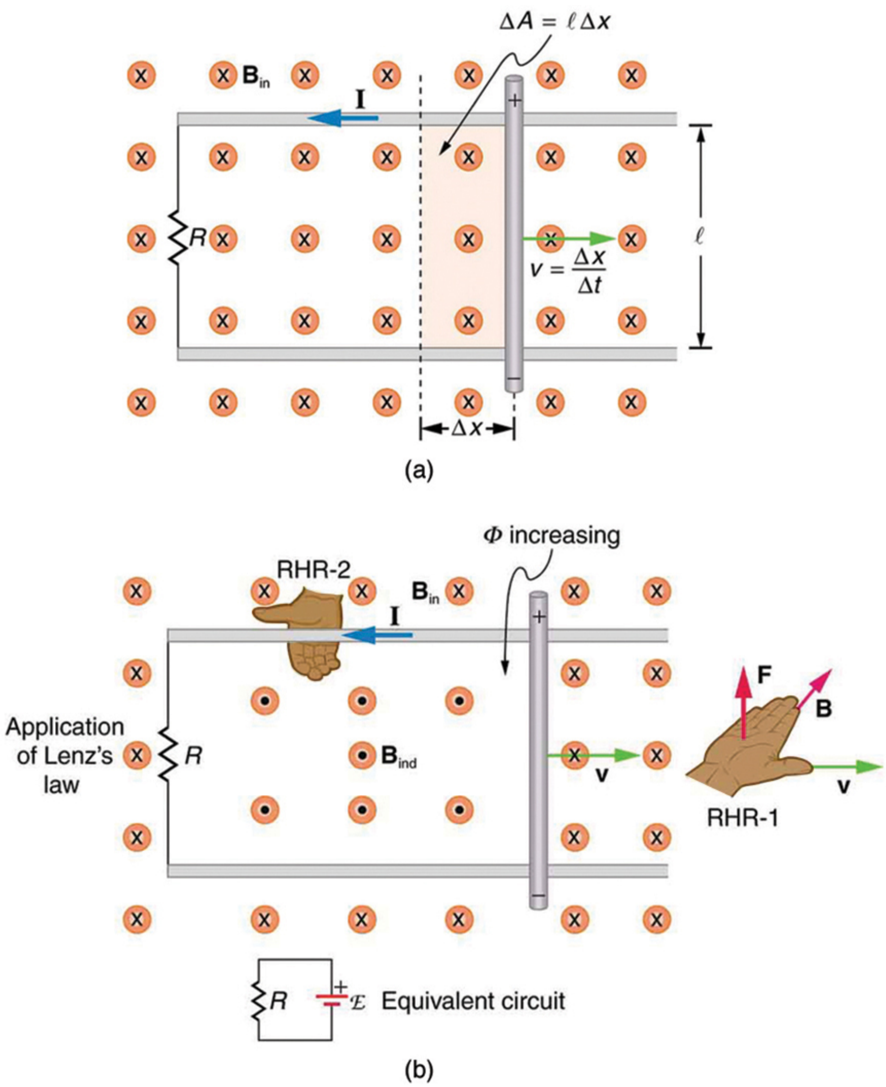
Motional Emf

- Calculate emf, force, magnetic field, and work due to the motion of an object in a magnetic field.

As we have seen, any change in magnetic flux induces an emf opposing that change—a process known as induction. Motion is one of the major causes of induction. For example, a magnet moved toward a coil induces an emf, and a coil moved toward a magnet produces a similar emf. In this section, we concentrate on motion in a magnetic field that is stationary relative to the Earth, producing what is loosely called *motional emf*.

One situation where motional emf occurs is known as the Hall effect and has already been examined. Charges moving in a magnetic field experience the magnetic force $F = \text{qvB} \sin \theta$, which moves opposite charges in opposite directions and produces an $\text{emf} = B\ell v$. We saw that the Hall effect has applications, including measurements of B and v. We will now see that the Hall effect is one aspect of the broader phenomenon of induction, and we will find that motional emf can be used as a power source.

Consider the situation shown in [link]. A rod is moved at a speed v along a pair of conducting rails separated by a distance ℓ in a uniform magnetic field B. The rails are stationary relative to B and are connected to a stationary resistor R. The resistor could be anything from a light bulb to a voltmeter. Consider the area enclosed by the moving rod, rails, and resistor. B is perpendicular to this area, and the area is increasing as the rod moves. Thus the magnetic flux enclosed by the rails, rod, and resistor is increasing. When flux changes, an emf is induced according to Faraday’s law of induction.



(a) A motional emf $= B\ell v$ is induced between the rails when this rod moves to the right in the uniform magnetic field. The magnetic field B is into the page, perpendicular to the moving rod and rails and, hence, to the area enclosed by them. (b) Lenz's law gives the directions of the induced field and current, and the polarity of the induced emf. Since the flux is increasing, the induced field is in the opposite direction, or out of the page. RHR-2 gives the current direction shown, and the polarity of the rod will drive such a current. RHR-1 also indicates the same polarity for the rod. (Note that

the script E symbol used in the equivalent circuit
at the bottom of part (b) represents emf.)

To find the magnitude of emf induced along the moving rod, we use Faraday's law of induction without the sign:

Equation:

$$\text{emf} = N\frac{\Delta\Phi}{\Delta t}.$$

Here and below, "emf" implies the magnitude of the emf. In this equation, $N = 1$ and the flux $\Phi = \text{BA}\cos\theta$. We have $\theta = 0^{\text{o}}$ and $\cos\theta = 1$, since B is perpendicular to A. Now $\Delta\Phi = \Delta(\text{BA}) = B\Delta A$, since B is uniform. Note that the area swept out by the rod is $\Delta A = \ell\Delta x$. Entering these quantities into the expression for emf yields

Equation:

$$\text{emf} = \frac{B\Delta A}{\Delta t} = B\frac{\ell\Delta x}{\Delta t}.$$

Finally, note that $\Delta x/\Delta t = v$, the velocity of the rod. Entering this into the last expression shows that

Equation:

$$\text{emf} = B\ell v \qquad (B,\ \ell,\ \text{and } v \text{ perpendicular})$$

is the motional emf. This is the same expression given for the Hall effect previously.

Note:
Making Connections: Unification of Forces

There are many connections between the electric force and the magnetic force. The fact that a moving electric field produces a magnetic field and, conversely, a moving magnetic field produces an electric field is part of why electric and magnetic forces are now considered to be different manifestations of the same force. This classic unification of electric and magnetic forces into what is called the electromagnetic force is the inspiration for contemporary efforts to unify other basic forces.

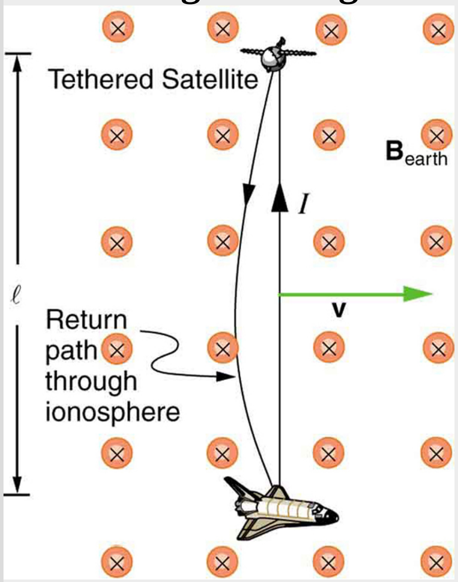
To find the direction of the induced field, the direction of the current, and the polarity of the induced emf, we apply Lenz’s law as explained in Faraday's Law of Induction: Lenz's Law. (See [link](b).) Flux is increasing, since the area enclosed is increasing. Thus the induced field must oppose the existing one and be out of the page. And so the RHR-2 requires that *I* be counterclockwise, which in turn means the top of the rod is positive as shown.

Motional emf also occurs if the magnetic field moves and the rod (or other object) is stationary relative to the Earth (or some observer). We have seen an example of this in the situation where a moving magnet induces an emf in a stationary coil. It is the relative motion that is important. What is emerging in these observations is a connection between magnetic and electric fields. A moving magnetic field produces an electric field through its induced emf. We already have seen that a moving electric field produces a magnetic field—moving charge implies moving electric field and moving charge produces a magnetic field.

Motional emfs in the Earth’s weak magnetic field are not ordinarily very large, or we would notice voltage along metal rods, such as a screwdriver, during ordinary motions. For example, a simple calculation of the motional emf of a 1 m rod moving at 3.0 m/s perpendicular to the Earth’s field gives $\text{emf} = B\ell v = (5.0 \times 10^{-5}\ \text{T})(1.0\ \text{m})(3.0\ \text{m/s}) = 150\ \mu\text{V}$. This small value is consistent with experience. There is a spectacular exception, however. In 1992 and 1996, attempts were made with the space shuttle to create large motional emfs. The Tethered Satellite was to be let out on a 20 km length of wire as shown in [link], to create a 5 kV emf by moving at

orbital speed through the Earth's field. This emf could be used to convert some of the shuttle's kinetic and potential energy into electrical energy if a complete circuit could be made. To complete the circuit, the stationary ionosphere was to supply a return path for the current to flow. (The ionosphere is the rarefied and partially ionized atmosphere at orbital altitudes. It conducts because of the ionization. The ionosphere serves the same function as the stationary rails and connecting resistor in [link], without which there would not be a complete circuit.) Drag on the current in the cable due to the magnetic force $F = I\ell B \sin \theta$ does the work that reduces the shuttle's kinetic and potential energy and allows it to be converted to electrical energy. The tests were both unsuccessful. In the first, the cable hung up and could only be extended a couple of hundred meters; in the second, the cable broke when almost fully extended. [link] indicates feasibility in principle.

Example:
Calculating the Large Motional Emf of an Object in Orbit



Motional emf as electrical power conversion for the space shuttle is the motivation for the

Tethered Satellite experiment. A 5 kV emf was predicted to be induced in the 20 km long tether while moving at orbital speed in the Earth's magnetic field. The circuit is completed by a return path through the stationary ionosphere.

Calculate the motional emf induced along a 20.0 km long conductor moving at an orbital speed of 7.80 km/s perpendicular to the Earth's 5.00×10^{-5} T magnetic field.

Strategy

This is a straightforward application of the expression for motional emf—emf $= B\ell v$.

Solution

Entering the given values into emf $= B\ell v$ gives

Equation:

$$\begin{aligned} \text{emf} &= B\ell v \\ &= (5.00 \times 10^{-5}\ \text{T})(2.0 \times 10^{4}\ \text{m})(7.80 \times 10^{3}\ \text{m/s}) \\ &= 7.80 \times 10^{3}\ \text{V}. \end{aligned}$$

Discussion

The value obtained is greater than the 5 kV measured voltage for the shuttle experiment, since the actual orbital motion of the tether is not perpendicular to the Earth's field. The 7.80 kV value is the maximum emf obtained when $\theta = 90^{\circ}$ and $\sin\theta = 1$.

Section Summary

- An emf induced by motion relative to a magnetic field B is called a *motional emf* and is given by
  Equation:
  $$\text{emf} = B\ell v \qquad (B,\, \ell,\, \text{and } v \text{ perpendicular}),$$
  where ℓ is the length of the object moving at speed v relative to the field.

Conceptual Questions

Exercise:

Problem:

Why must part of the circuit be moving relative to other parts, to have usable motional emf? Consider, for example, that the rails in [link] are stationary relative to the magnetic field, while the rod moves.

Exercise:

Problem:

A powerful induction cannon can be made by placing a metal cylinder inside a solenoid coil. The cylinder is forcefully expelled when solenoid current is turned on rapidly. Use Faraday’s and Lenz’s laws to explain how this works. Why might the cylinder get live/hot when the cannon is fired?

Exercise:

Problem:

An induction stove heats a pot with a coil carrying an alternating current located beneath the pot (and without a hot surface). Can the stove surface be a conductor? Why won’t a coil carrying a direct current work?

Exercise:

Problem:

Explain how you could thaw out a frozen water pipe by wrapping a coil carrying an alternating current around it. Does it matter whether or not the pipe is a conductor? Explain.

Problems & Exercises

Exercise:

Problem:

Use Faraday's law, Lenz's law, and RHR-1 to show that the magnetic force on the current in the moving rod in [link] is in the opposite direction of its velocity.

Exercise:

Problem:

If a current flows in the Satellite Tether shown in [link], use Faraday's law, Lenz's law, and RHR-1 to show that there is a magnetic force on the tether in the direction opposite to its velocity.

Exercise:

Problem:

(a) A jet airplane with a 75.0 m wingspan is flying at 280 m/s. What emf is induced between wing tips if the vertical component of the Earth's field is $3.00 \times 10^{-5}\ \mathrm{T}$? (b) Is an emf of this magnitude likely to have any consequences? Explain.

Solution:

(a) 0.630 V

(b) No, this is a very small emf.

Exercise:

Problem:

(a) A nonferrous screwdriver is being used in a 2.00 T magnetic field. What maximum emf can be induced along its 12.0 cm length when it moves at 6.00 m/s? (b) Is it likely that this emf will have any consequences or even be noticed?

Exercise:

Problem:

At what speed must the sliding rod in [link] move to produce an emf of 1.00 V in a 1.50 T field, given the rod's length is 30.0 cm?

Solution:

2.22 m/s

Exercise:

Problem:

The 12.0 cm long rod in [link] moves at 4.00 m/s. What is the strength of the magnetic field if a 95.0 V emf is induced?

Exercise:

Problem:

Prove that when B, ℓ, and v are not mutually perpendicular, motional emf is given by emf $= B\ell v\sin\theta$. If v is perpendicular to B, then θ is the angle between ℓ and B. If ℓ is perpendicular to B, then θ is the angle between v and B.

Exercise:

Problem:

In the August 1992 space shuttle flight, only 250 m of the conducting tether considered in [link] could be let out. A 40.0 V motional emf was generated in the Earth's $5.00 \times 10^{-5}\ \text{T}$ field, while moving at $7.80 \times 10^{3}\ \text{m/s}$. What was the angle between the shuttle's velocity and the Earth's field, assuming the conductor was perpendicular to the field?

Exercise:

Problem: **Integrated Concepts**

Derive an expression for the current in a system like that in [link], under the following conditions. The resistance between the rails is R, the rails and the moving rod are identical in cross section A and have the same resistivity ρ. The distance between the rails is l, and the rod moves at constant speed v perpendicular to the uniform field B. At time zero, the moving rod is next to the resistance R.

Exercise:

Problem: **Integrated Concepts**

The Tethered Satellite in [link] has a mass of 525 kg and is at the end of a 20.0 km long, 2.50 mm diameter cable with the tensile strength of steel. (a) How much does the cable stretch if a 100 N force is exerted to pull the satellite in? (Assume the satellite and shuttle are at the same altitude above the Earth.) (b) What is the effective force constant of the cable? (c) How much energy is stored in it when stretched by the 100 N force?

Exercise:

Problem: **Integrated Concepts**

The Tethered Satellite discussed in this module is producing 5.00 kV, and a current of 10.0 A flows. (a) What magnetic drag force does this

produce if the system is moving at 7.80 km/s? (b) How much kinetic energy is removed from the system in 1.00 h, neglecting any change in altitude or velocity during that time? (c) What is the change in velocity if the mass of the system is 100,000 kg? (d) Discuss the long term consequences (say, a week-long mission) on the space shuttle's orbit, noting what effect a decrease in velocity has and assessing the magnitude of the effect.

Solution:

(a) 10.0 N

(b) 2.81×10^8 J

(c) 0.36 m/s

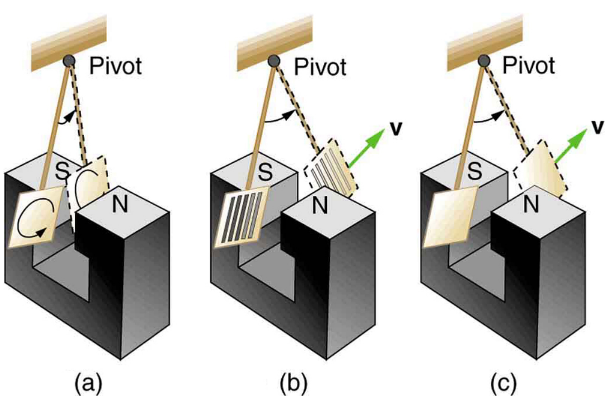
(d) For a week-long mission (168 hours), the change in velocity will be 60 m/s, or approximately 1%. In general, a decrease in velocity would cause the orbit to start spiraling inward because the velocity would no longer be sufficient to keep the circular orbit. The long-term consequences are that the shuttle would require a little more fuel to maintain the desired speed, otherwise the orbit would spiral slightly inward.

Eddy Currents and Magnetic Damping

- Explain the magnitude and direction of an induced eddy current, and the effect this will have on the object it is induced in.
- Describe several applications of magnetic damping.

Eddy Currents and Magnetic Damping

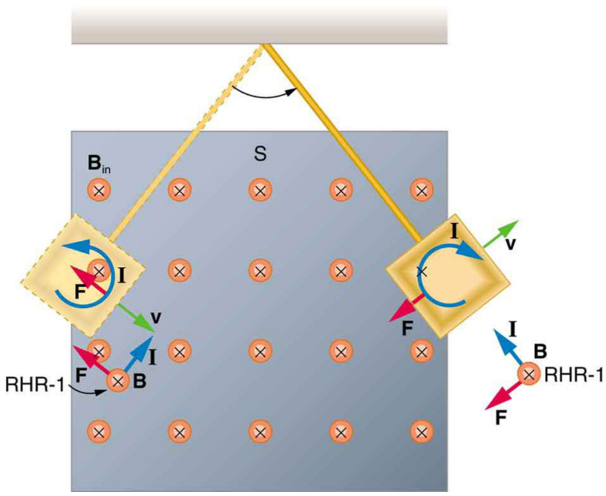
As discussed in Motional Emf, motional emf is induced when a conductor moves in a magnetic field or when a magnetic field moves relative to a conductor. If motional emf can cause a current loop in the conductor, we refer to that current as an **eddy current**. Eddy currents can produce significant drag, called **magnetic damping**, on the motion involved. Consider the apparatus shown in [link], which swings a pendulum bob between the poles of a strong magnet. (This is another favorite physics lab activity.) If the bob is metal, there is significant drag on the bob as it enters and leaves the field, quickly damping the motion. If, however, the bob is a slotted metal plate, as shown in [link](b), there is a much smaller effect due to the magnet. There is no discernible effect on a bob made of an insulator. Why is there drag in both directions, and are there any uses for magnetic drag?



A common physics demonstration device for exploring eddy currents and magnetic damping. (a) The

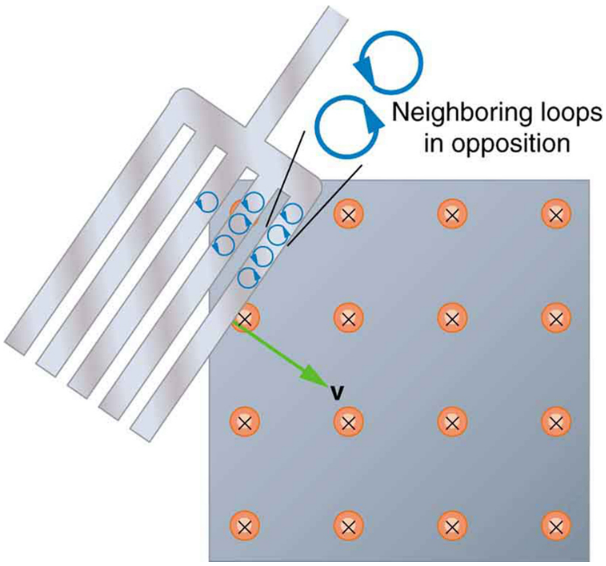
motion of a metal pendulum bob swinging between the poles of a magnet is quickly damped by the action of eddy currents. (b) There is little effect on the motion of a slotted metal bob, implying that eddy currents are made less effective. (c) There is also no magnetic damping on a nonconducting bob, since the eddy currents are extremely small.

[link] shows what happens to the metal plate as it enters and leaves the magnetic field. In both cases, it experiences a force opposing its motion. As it enters from the left, flux increases, and so an eddy current is set up (Faraday's law) in the counterclockwise direction (Lenz's law), as shown. Only the right-hand side of the current loop is in the field, so that there is an unopposed force on it to the left (RHR-1). When the metal plate is completely inside the field, there is no eddy current if the field is uniform, since the flux remains constant in this region. But when the plate leaves the field on the right, flux decreases, causing an eddy current in the clockwise direction that, again, experiences a force to the left, further slowing the motion. A similar analysis of what happens when the plate swings from the right toward the left shows that its motion is also damped when entering and leaving the field.



A more detailed look at the conducting plate passing between the poles of a magnet. As it enters and leaves the field, the change in flux produces an eddy current. Magnetic force on the current loop opposes the motion. There is no current and no magnetic drag when the plate is completely inside the uniform field.

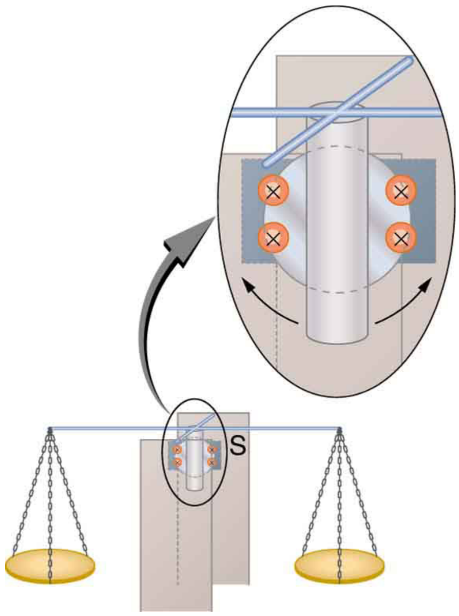
When a slotted metal plate enters the field, as shown in [link], an emf is induced by the change in flux, but it is less effective because the slots limit the size of the current loops. Moreover, adjacent loops have currents in opposite directions, and their effects cancel. When an insulating material is used, the eddy current is extremely small, and so magnetic damping on insulators is negligible. If eddy currents are to be avoided in conductors, then they can be slotted or constructed of thin layers of conducting material separated by insulating sheets.



Eddy currents induced in a slotted metal plate entering a magnetic field form small loops, and the forces on them tend to cancel, thereby making magnetic drag almost zero.

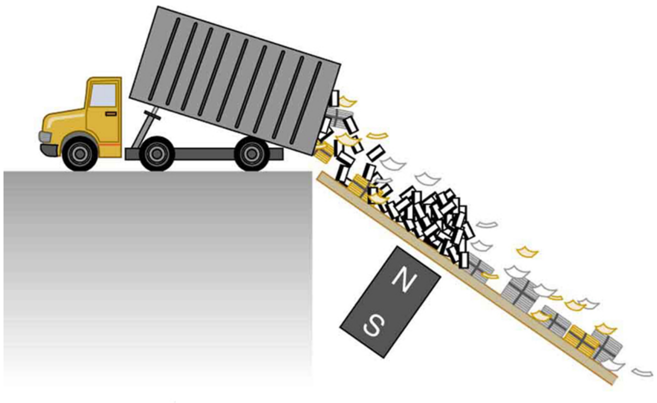
Applications of Magnetic Damping

One use of magnetic damping is found in sensitive laboratory balances. To have maximum sensitivity and accuracy, the balance must be as friction-free as possible. But if it is friction-free, then it will oscillate for a very long time. Magnetic damping is a simple and ideal solution. With magnetic damping, drag is proportional to speed and becomes zero at zero velocity. Thus the oscillations are quickly damped, after which the damping force disappears, allowing the balance to be very sensitive. (See [link].) In most balances, magnetic damping is accomplished with a conducting disc that rotates in a fixed field.



Magnetic damping of this sensitive balance slows its oscillations. Since Faraday’s law of induction gives the greatest effect for the most rapid change, damping is greatest for large oscillations and goes to zero as the motion stops.

Since eddy currents and magnetic damping occur only in conductors, recycling centers can use magnets to separate metals from other materials. Trash is dumped in batches down a ramp, beneath which lies a powerful magnet. Conductors in the trash are slowed by magnetic damping while nonmetals in the trash move on, separating from the metals. (See [link].) This works for all metals, not just ferromagnetic ones. A magnet can separate out the ferromagnetic materials alone by acting on stationary trash.



Metals can be separated from other trash by magnetic drag. Eddy currents and magnetic drag are created in the metals sent down this ramp by the powerful magnet beneath it. Nonmetals move on.

Other major applications of eddy currents are in metal detectors and braking systems in trains and roller coasters. Portable metal detectors ([link]) consist of a primary coil carrying an alternating current and a secondary coil in which a current is induced. An eddy current will be induced in a piece of metal close to the detector which will cause a change in the induced current within the secondary coil, leading to some sort of signal like a shrill noise. Braking using eddy currents is safer because factors such as rain do not affect the braking and the braking is smoother. However, eddy currents cannot bring the motion to a complete stop, since the force produced decreases with speed. Thus, speed can be reduced from say 20 m/s to 5 m/s, but another form of braking is needed to completely stop the vehicle. Generally, powerful rare earth magnets such as neodymium magnets are used in roller coasters. [link] shows rows of magnets in such an application. The vehicle has metal fins (normally containing copper) which pass through the magnetic field slowing the vehicle down in much the same way as with the pendulum bob shown in [link].

A soldier in Iraq uses a metal detector to search for explosives and weapons. (credit: U.S. Army)

The rows of rare earth magnets (protruding horizontally) are used for magnetic braking in roller coasters. (credit: Stefan Scheer, Wikimedia Commons)

Induction cooktops have electromagnets under their surface. The magnetic field is varied rapidly producing eddy currents in the base of the pot, causing the pot and its contents to increase in temperature. Induction cooktops have high efficiencies and good response times but the base of the pot needs to be ferromagnetic, iron or steel for induction to work.

Section Summary

- Current loops induced in moving conductors are called eddy currents.
- They can create significant drag, called magnetic damping.

Conceptual Questions

Exercise:

Problem:

Explain why magnetic damping might not be effective on an object made of several thin conducting layers separated by insulation.

Exercise:

Problem:

Explain how electromagnetic induction can be used to detect metals? This technique is particularly important in detecting buried landmines for disposal, geophysical prospecting and at airports.

Problems & Exercises

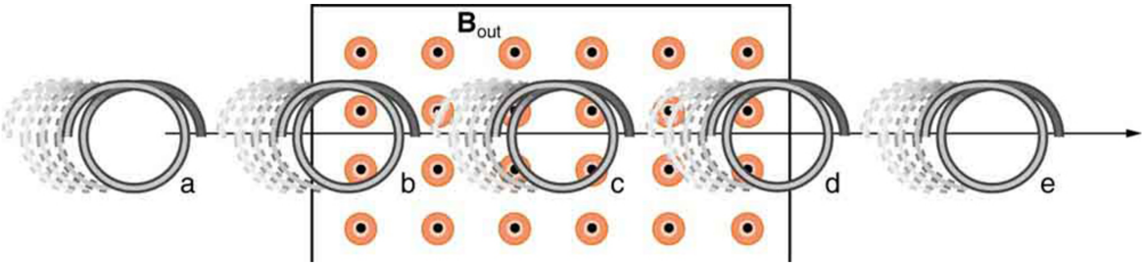
Exercise:

Problem:

Make a drawing similar to [link], but with the pendulum moving in the opposite direction. Then use Faraday’s law, Lenz’s law, and RHR-1 to show that magnetic force opposes motion.

Exercise:

Problem:



A coil is moved into and out of a region of uniform magnetic field.

A coil is moved through a magnetic field as shown in [link]. The field is uniform inside the rectangle and zero outside. What is the direction of the induced current and what is the direction of the magnetic force on the coil at each position shown?

Glossary

eddy current
: a current loop in a conductor caused by motional emf

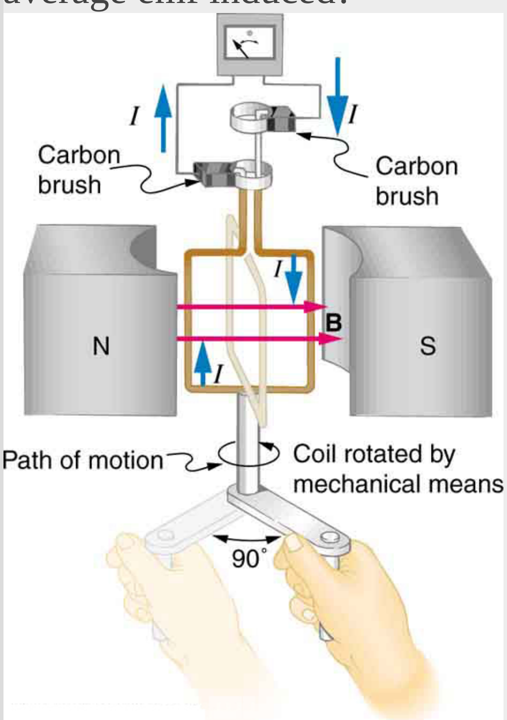
magnetic damping
: the drag produced by eddy currents

Electric Generators

- Calculate the emf induced in a generator.
- Calculate the peak emf which can be induced in a particular generator system.

Electric generators induce an emf by rotating a coil in a magnetic field, as briefly discussed in Induced Emf and Magnetic Flux. We will now explore generators in more detail. Consider the following example.

Example:
Calculating the Emf Induced in a Generator Coil
The generator coil shown in [link] is rotated through one-fourth of a revolution (from $\theta = 0^{\text{o}}$ to $\theta = 90^{\text{o}}$) in 15.0 ms. The 200-turn circular coil has a 5.00 cm radius and is in a uniform 1.25 T magnetic field. What is the average emf induced?



When this generator coil is rotated through one-fourth of a revolution, the

magnetic flux Φ changes from its maximum to zero, inducing an emf.

Strategy

We use Faraday's law of induction to find the average emf induced over a time Δt:

Equation:

$$\text{emf} = -N\frac{\Delta\Phi}{\Delta t}.$$

We know that $N = 200$ and $\Delta t = 15.0\text{ ms}$, and so we must determine the change in flux $\Delta\Phi$ to find emf.

Solution

Since the area of the loop and the magnetic field strength are constant, we see that

Equation:

$$\Delta\Phi = \Delta(BA\cos\theta) = AB\Delta(\cos\theta).$$

Now, $\Delta(\cos\theta) = -1.0$, since it was given that θ goes from 0º to 90º . Thus $\Delta\Phi = -\text{AB}$, and

Equation:

$$\text{emf} = N\frac{\text{AB}}{\Delta t}.$$

The area of the loop is $A = \pi r^2 = (3.14...)(0.0500\text{ m})^2 = 7.85 \times 10^{-3}\text{ m}^2$. Entering this value gives

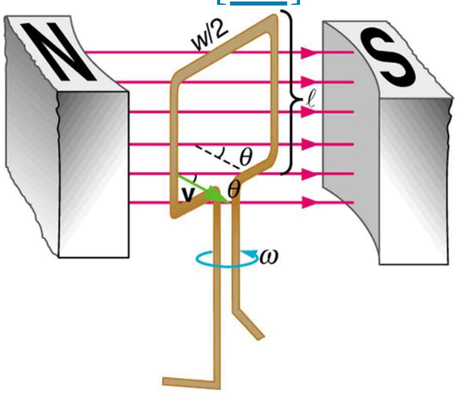
Equation:

$$\text{emf} = 200\frac{(7.85 \times 10^{-3}\text{ m}^2)(1.25\text{ T})}{15.0 \times 10^{-3}\text{ s}} = 131\text{ V}.$$

Discussion

This is a practical average value, similar to the 120 V used in household power.

The emf calculated in [link] is the average over one-fourth of a revolution. What is the emf at any given instant? It varies with the angle between the magnetic field and a perpendicular to the coil. We can get an expression for emf as a function of time by considering the motional emf on a rotating rectangular coil of width w and height ℓ in a uniform magnetic field, as illustrated in [link].



A generator with a single rectangular coil rotated at constant angular velocity in a uniform magnetic field produces an emf that varies sinusoidally in time. Note the generator is similar to a motor, except the shaft is rotated to produce a current rather than the other way around.

Charges in the wires of the loop experience the magnetic force, because they are moving in a magnetic field. Charges in the vertical wires experience forces parallel to the wire, causing currents. But those in the top and bottom segments feel a force perpendicular to the wire, which does not cause a current. We can thus find the induced emf by considering only the side wires. Motional emf is given to be $\text{emf} = B\ell v$, where the velocity v is perpendicular to the magnetic field B. Here the velocity is at an angle θ with B, so that its component perpendicular to B is $v \sin \theta$ (see [link]). Thus in this case the emf induced on each side is $\text{emf} = B\ell v \sin \theta$, and they are in the same direction. The total emf around the loop is then

Equation:

$$\text{emf} = 2B\ell v \sin \theta.$$

This expression is valid, but it does not give emf as a function of time. To find the time dependence of emf, we assume the coil rotates at a constant angular velocity ω. The angle θ is related to angular velocity by $\theta = \omega t$, so that

Equation:

$$\text{emf} = 2B\ell v \sin \omega t.$$

Now, linear velocity v is related to angular velocity ω by $v = r\omega$. Here $r = w/2$, so that $v = (w/2)\omega$, and

Equation:

$$\text{emf} = 2B\ell \frac{w}{2} \omega \sin \omega t = (\ell w) B\omega \sin \omega t.$$

Noting that the area of the loop is $A = \ell w$, and allowing for N loops, we find that

Equation:

$$\text{emf} = \text{NAB}\omega \sin \omega t$$

is the **emf induced in a generator coil** of N turns and area A rotating at a constant angular velocity ω in a uniform magnetic field B. This can also be expressed as

Equation:

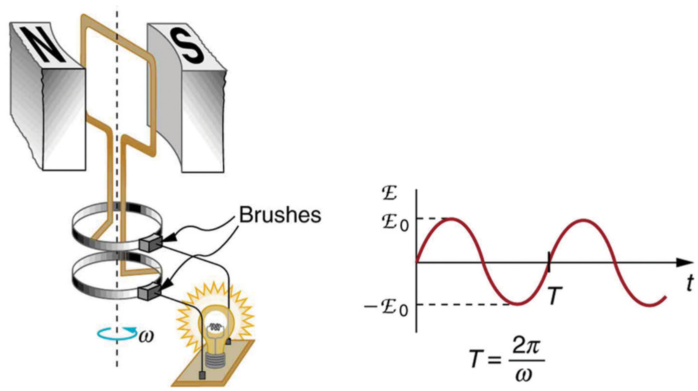
$$\text{emf} = \text{emf}_0 \sin \omega t,$$

where

Equation:

$$\text{emf}_0 = \text{NAB}\omega$$

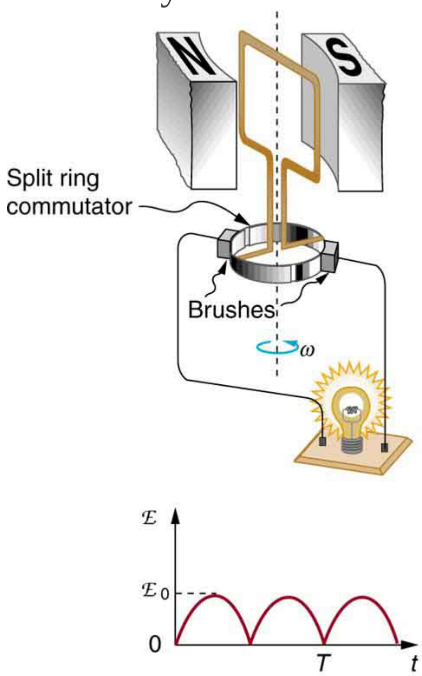
is the maximum **(peak) emf**. Note that the frequency of the oscillation is $f = \omega / 2\pi$, and the period is $T = 1/f = 2\pi/\omega$. [link] shows a graph of emf as a function of time, and it now seems reasonable that AC voltage is sinusoidal.



The emf of a generator is sent to a light bulb with the system of rings and brushes shown. The graph gives the emf of the generator as a function of time. emf_0 is the peak emf. The period is $T = 1/f = 2\pi/\omega$, where f is the frequency. Note that the script E stands for emf.

The fact that the peak emf, $\mathrm{emf}_0 = \mathrm{NAB}\omega$, makes good sense. The greater the number of coils, the larger their area, and the stronger the field, the greater the output voltage. It is interesting that the faster the generator is spun (greater ω), the greater the emf. This is noticeable on bicycle generators—at least the cheaper varieties. One of the authors as a juvenile found it amusing to ride his bicycle fast enough to burn out his lights, until he had to ride home lightless one dark night.

[link] shows a scheme by which a generator can be made to produce pulsed DC. More elaborate arrangements of multiple coils and split rings can produce smoother DC, although electronic rather than mechanical means are usually used to make ripple-free DC.



Split rings, called commutators, produce a pulsed DC emf output in this configuration.

Example:

Calculating the Maximum Emf of a Generator

Calculate the maximum emf, emf_0, of the generator that was the subject of [link].

Strategy

Once ω, the angular velocity, is determined, $\text{emf}_0 = \text{NAB}\omega$ can be used to find emf_0. All other quantities are known.

Solution

Angular velocity is defined to be the change in angle per unit time:

Equation:

$$\omega = \frac{\Delta\theta}{\Delta t}.$$

One-fourth of a revolution is $\pi/2$ radians, and the time is 0.0150 s; thus,

Equation:

$$\begin{aligned} \omega &= \frac{\pi/2 \text{ rad}}{0.0150 \text{ s}} \\ &= 104.7 \text{ rad/s}. \end{aligned}$$

104.7 rad/s is exactly 1000 rpm. We substitute this value for ω and the information from the previous example into $\text{emf}_0 = \text{NAB}\omega$, yielding

Equation:

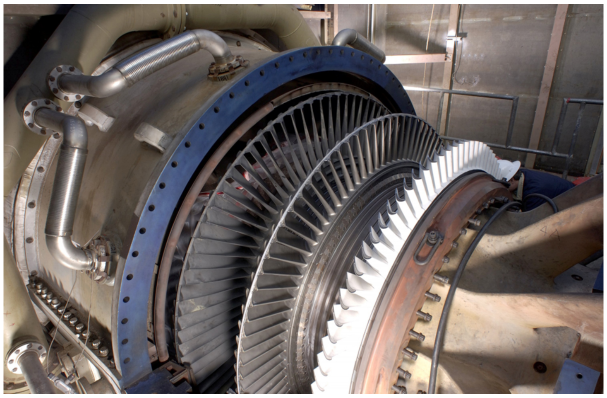
$$\begin{aligned} \text{emf}_0 &= NAB\omega \\ &= 200(7.85 \times 10^{-3} \text{ m}^2)(1.25 \text{ T})(104.7 \text{ rad/s}). \\ &= 206 \text{ V} \end{aligned}$$

Discussion

The maximum emf is greater than the average emf of 131 V found in the previous example, as it should be.

In real life, electric generators look a lot different than the figures in this section, but the principles are the same. The source of mechanical energy that turns the coil can be falling water (hydropower), steam produced by the

burning of fossil fuels, or the kinetic energy of wind. [link] shows a cutaway view of a steam turbine; steam moves over the blades connected to the shaft, which rotates the coil within the generator.



Steam turbine/generator. The steam produced by burning coal impacts the turbine blades, turning the shaft which is connected to the generator. (credit: Nabonaco, Wikimedia Commons)

Generators illustrated in this section look very much like the motors illustrated previously. This is not coincidental. In fact, a motor becomes a generator when its shaft rotates. Certain early automobiles used their starter motor as a generator. In Back Emf, we shall further explore the action of a motor as a generator.

Section Summary

- An electric generator rotates a coil in a magnetic field, inducing an emfgiven as a function of time by
  Equation:

$$\text{emf} = \text{NAB}\omega \sin \omega t,$$

where A is the area of an N-turn coil rotated at a constant angular velocity ω in a uniform magnetic field B.
- The peak emf emf_0 of a generator is
  Equation:

$$\text{emf}_0 = \text{NAB}\omega.$$

Conceptual Questions

Exercise:

Problem:

Using RHR-1, show that the emfs in the sides of the generator loop in [link] are in the same sense and thus add.

Exercise:

Problem:

The source of a generator's electrical energy output is the work done to turn its coils. How is the work needed to turn the generator related to Lenz's law?

Problems & Exercises

Exercise:

Problem:

Calculate the peak voltage of a generator that rotates its 200-turn, 0.100 m diameter coil at 3600 rpm in a 0.800 T field.

Solution:

474 V

Exercise:

Problem:

At what angular velocity in rpm will the peak voltage of a generator be 480 V, if its 500-turn, 8.00 cm diameter coil rotates in a 0.250 T field?

Exercise:

Problem:

What is the peak emf generated by rotating a 1000-turn, 20.0 cm diameter coil in the Earth's $5.00 \times 10^{-5}\ \mathrm{T}$ magnetic field, given the plane of the coil is originally perpendicular to the Earth's field and is rotated to be parallel to the field in 10.0 ms?

Solution:

0.247 V

Exercise:

Problem:

What is the peak emf generated by a 0.250 m radius, 500-turn coil is rotated one-fourth of a revolution in 4.17 ms, originally having its plane perpendicular to a uniform magnetic field. (This is 60 rev/s.)

Exercise:

Problem:

(a) A bicycle generator rotates at 1875 rad/s, producing an 18.0 V peak emf. It has a 1.00 by 3.00 cm rectangular coil in a 0.640 T field. How many turns are in the coil? (b) Is this number of turns of wire practical for a 1.00 by 3.00 cm coil?

Solution:

(a) 50

(b) yes

Exercise:

Problem: Integrated Concepts

This problem refers to the bicycle generator considered in the previous problem. It is driven by a 1.60 cm diameter wheel that rolls on the outside rim of the bicycle tire. (a) What is the velocity of the bicycle if the generator's angular velocity is 1875 rad/s? (b) What is the maximum emf of the generator when the bicycle moves at 10.0 m/s, noting that it was 18.0 V under the original conditions? (c) If the sophisticated generator can vary its own magnetic field, what field strength will it need at 5.00 m/s to produce a 9.00 V maximum emf?

Exercise:

Problem:

(a) A car generator turns at 400 rpm when the engine is idling. Its 300-turn, 5.00 by 8.00 cm rectangular coil rotates in an adjustable magnetic field so that it can produce sufficient voltage even at low rpms. What is the field strength needed to produce a 24.0 V peak emf? (b) Discuss how this required field strength compares to those available in permanent and electromagnets.

Solution:

(a) 0.477 T

(b) This field strength is small enough that it can be obtained using either a permanent magnet or an electromagnet.

Exercise:

Problem:

Show that if a coil rotates at an angular velocity ω, the period of its AC output is $2\pi/\omega$.

Exercise:

Problem:

A 75-turn, 10.0 cm diameter coil rotates at an angular velocity of 8.00 rad/s in a 1.25 T field, starting with the plane of the coil parallel to the field. (a) What is the peak emf? (b) At what time is the peak emf first reached? (c) At what time is the emf first at its most negative? (d) What is the period of the AC voltage output?

Solution:

(a) 5.89 V

(b) At t=0

(c) 0.393 s

(d) 0.785 s

Exercise:

Problem:

(a) If the emf of a coil rotating in a magnetic field is zero at $t = 0$, and increases to its first peak at $t = 0.100 \text{ ms}$, what is the angular velocity of the coil? (b) At what time will its next maximum occur? (c) What is the period of the output? (d) When is the output first one-fourth of its maximum? (e) When is it next one-fourth of its maximum?

Exercise:

Problem: Unreasonable Results

A 500-turn coil with a 0.250 m^2 area is spun in the Earth's $5.00 \times 10^{-5} \text{ T}$ field, producing a 12.0 kV maximum emf. (a) At what angular velocity must the coil be spun? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Solution:

(a) $1.92 \times 10^6 \text{ rad/s}$

(b) This angular velocity is unreasonably high, higher than can be obtained for any mechanical system.

(c) The assumption that a voltage as great as 12.0 kV could be obtained is unreasonable.

Glossary

electric generator
: a device for converting mechanical work into electric energy; it induces an emf by rotating a coil in a magnetic field

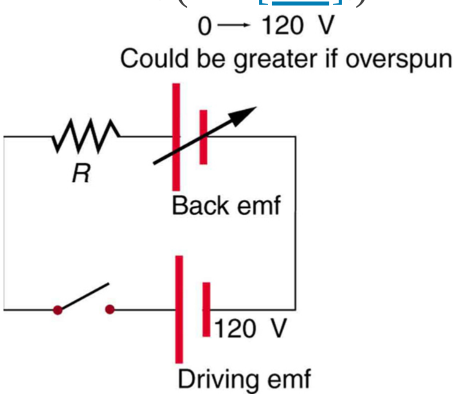
emf induced in a generator coil
: $\text{emf} = \text{NAB}\omega \sin \omega t$, where A is the area of an N-turn coil rotated at a constant angular velocity ω in a uniform magnetic field B, over a period of time t

peak emf
: $\text{emf}_0 = \text{NAB}\omega$

Back Emf

- Explain what back emf is and how it is induced.

It has been noted that motors and generators are very similar. Generators convert mechanical energy into electrical energy, whereas motors convert electrical energy into mechanical energy. Furthermore, motors and generators have the same construction. When the coil of a motor is turned, magnetic flux changes, and an emf (consistent with Faraday's law of induction) is induced. The motor thus acts as a generator whenever its coil rotates. This will happen whether the shaft is turned by an external input, like a belt drive, or by the action of the motor itself. That is, when a motor is doing work and its shaft is turning, an emf is generated. Lenz's law tells us the emf opposes any change, so that the input emf that powers the motor will be opposed by the motor's self-generated emf, called the **back emf** of the motor. (See [link].)



The coil of a DC motor is represented as a resistor in this schematic. The back emf is represented as a variable emf that opposes the one driving the motor. Back emf is zero when the motor is not turning, and it increases

proportionally to the
motor’s angular
velocity.

Back emf is the generator output of a motor, and so it is proportional to the motor’s angular velocity ω. It is zero when the motor is first turned on, meaning that the coil receives the full driving voltage and the motor draws maximum current when it is on but not turning. As the motor turns faster and faster, the back emf grows, always opposing the driving emf, and reduces the voltage across the coil and the amount of current it draws. This effect is noticeable in a number of situations. When a vacuum cleaner, refrigerator, or washing machine is first turned on, lights in the same circuit dim briefly due to the IR drop produced in feeder lines by the large current drawn by the motor. When a motor first comes on, it draws more current than when it runs at its normal operating speed. When a mechanical load is placed on the motor, like an electric wheelchair going up a hill, the motor slows, the back emf drops, more current flows, and more work can be done. If the motor runs at too low a speed, the larger current can overheat it (via resistive power in the coil, $P = I^2R$), perhaps even burning it out. On the other hand, if there is no mechanical load on the motor, it will increase its angular velocity ω until the back emf is nearly equal to the driving emf. Then the motor uses only enough energy to overcome friction.

Consider, for example, the motor coils represented in [link]. The coils have a $0.400\ \Omega$ equivalent resistance and are driven by a 48.0 V emf. Shortly after being turned on, they draw a current $I = \text{V/R} = (48.0\text{ V })/\ (0.400\ \Omega) = 120\text{ A}$ and, thus, dissipate $P = I^2R = 5.76\text{ kW}$ of energy as heat transfer. Under normal operating conditions for this motor, suppose the back emf is 40.0 V. Then at operating speed, the total voltage across the coils is 8.0 V (48.0 V minus the 40.0 V back emf), and the current drawn is $I = \text{V/R} = (8.0\text{ V})\ /(\ 0.400\ \Omega) = 20\text{ A}$. Under normal load, then, the power dissipated is $P = \text{IV} = (20\text{ A})/(8.0\text{ V}) = 160\text{ W}$. The latter will not cause a problem for this motor, whereas the former 5.76 kW would burn out the coils if sustained.

Section Summary

- Any rotating coil will have an induced emf—in motors, this is called back emf, since it opposes the emf input to the motor.

Conceptual Questions

Exercise:

Problem:

Suppose you find that the belt drive connecting a powerful motor to an air conditioning unit is broken and the motor is running freely. Should you be worried that the motor is consuming a great deal of energy for no useful purpose? Explain why or why not.

Problems & Exercises

Exercise:

Problem:

Suppose a motor connected to a 120 V source draws 10.0 A when it first starts. (a) What is its resistance? (b) What current does it draw at its normal operating speed when it develops a 100 V back emf?

Solution:

(a) $12.00\ \Omega$

(b) 1.67 A

Exercise:

Problem:

A motor operating on 240 V electricity has a 180 V back emf at operating speed and draws a 12.0 A current. (a) What is its resistance? (b) What current does it draw when it is first started?

Exercise:

Problem:

What is the back emf of a 120 V motor that draws 8.00 A at its normal speed and 20.0 A when first starting?

Solution:

72.0 V

Exercise:

Problem:

The motor in a toy car operates on 6.00 V, developing a 4.50 V back emf at normal speed. If it draws 3.00 A at normal speed, what current does it draw when starting?

Exercise:

Problem: **Integrated Concepts**

The motor in a toy car is powered by four batteries in series, which produce a total emf of 6.00 V. The motor draws 3.00 A and develops a 4.50 V back emf at normal speed. Each battery has a $0.100\ \Omega$ internal resistance. What is the resistance of the motor?

Solution:

$0.100\ \Omega$

Glossary

back emf
: the emf generated by a running motor, because it consists of a coil turning in a magnetic field; it opposes the voltage powering the motor

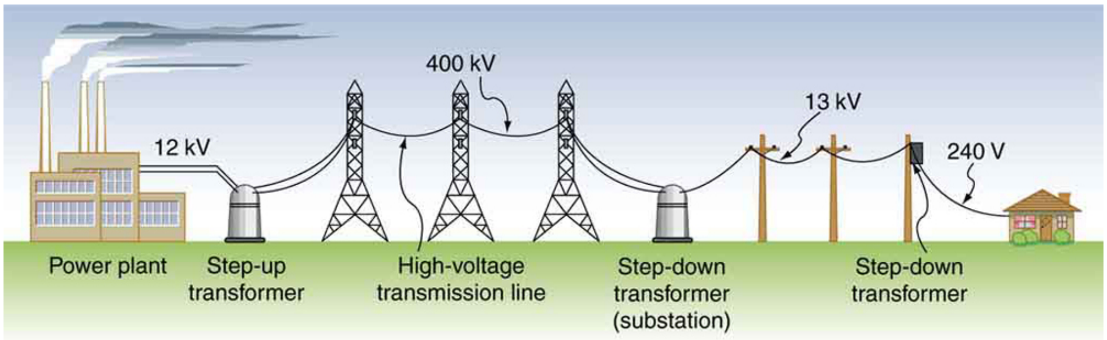
Transformers

- Explain how a transformer works.
- Calculate voltage, current, and/or number of turns given the other quantities.

Transformers do what their name implies—they transform voltages from one value to another (The term voltage is used rather than emf, because transformers have internal resistance). For example, many cell phones, laptops, video games, and power tools and small appliances have a transformer built into their plug-in unit (like that in [link]) that changes 120 V or 240 V AC into whatever voltage the device uses. Transformers are also used at several points in the power distribution systems, such as illustrated in [link]. Power is sent long distances at high voltages, because less current is required for a given amount of power, and this means less line loss, as was discussed previously. But high voltages pose greater hazards, so that transformers are employed to produce lower voltage at the user's location.

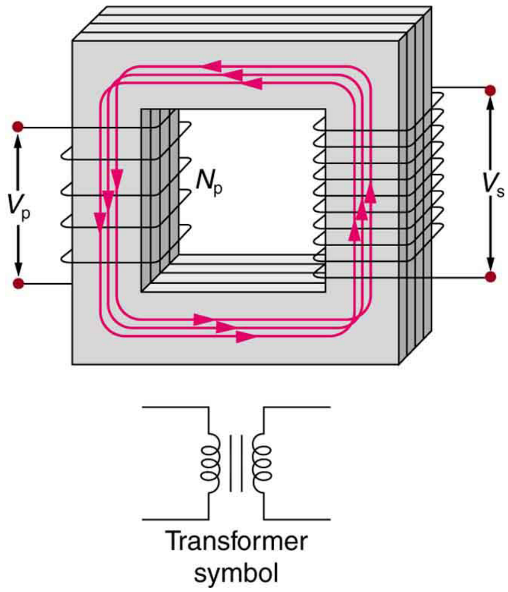
The plug-in transformer has become increasingly familiar with the proliferation of electronic devices that operate on voltages other than common 120 V

AC. Most are in the 3 to 12 V range. (credit: Shop Xtreme)



Transformers change voltages at several points in a power distribution system. Electric power is usually generated at greater than 10 kV, and transmitted long distances at voltages over 200 kV—sometimes as great as 700 kV—to limit energy losses. Local power distribution to neighborhoods or industries goes through a substation and is sent short distances at voltages ranging from 5 to 13 kV. This is reduced to 120, 240, or 480 V for safety at the individual user site.

The type of transformer considered in this text—see [link]—is based on Faraday's law of induction and is very similar in construction to the apparatus Faraday used to demonstrate magnetic fields could cause currents. The two coils are called the *primary* and *secondary coils*. In normal use, the input voltage is placed on the primary, and the secondary produces the transformed output voltage. Not only does the iron core trap the magnetic field created by the primary coil, its magnetization increases the field strength. Since the input voltage is AC, a time-varying magnetic flux is sent to the secondary, inducing its AC output voltage.



A typical construction of a simple transformer has two coils wound on a ferromagnetic core that is laminated to minimize eddy currents. The magnetic field created by the primary is mostly confined to and increased by the core, which transmits it to the secondary coil. Any change in current in the primary induces a current in the secondary.

For the simple transformer shown in [link], the output voltage V_s depends almost entirely on the input voltage V_p and the ratio of the number of loops in the primary and secondary coils. Faraday's law of induction for the secondary coil gives its induced output voltage V_s to be

Equation:

$$V_{\mathrm{s}} = -N_{\mathrm{s}} \frac{\Delta \Phi}{\Delta t},$$

where N_{s} is the number of loops in the secondary coil and $\Delta\Phi/\Delta t$ is the rate of change of magnetic flux. Note that the output voltage equals the induced emf ($V_{\mathrm{s}} = \mathrm{emf}_{\mathrm{s}}$), provided coil resistance is small (a reasonable assumption for transformers). The cross-sectional area of the coils is the same on either side, as is the magnetic field strength, and so $\Delta\Phi/\Delta t$ is the same on either side. The input primary voltage V_{p} is also related to changing flux by

Equation:

$$V_p = -N_{\mathrm{p}} \frac{\Delta \Phi}{\Delta t}.$$

The reason for this is a little more subtle. Lenz's law tells us that the primary coil opposes the change in flux caused by the input voltage V_{p}, hence the minus sign (This is an example of *self-inductance*, a topic to be explored in some detail in later sections). Assuming negligible coil resistance, Kirchhoff's loop rule tells us that the induced emf exactly equals the input voltage. Taking the ratio of these last two equations yields a useful relationship:

Equation:

$$\frac{V_{\mathrm{s}}}{V_{\mathrm{p}}} = \frac{N_{\mathrm{s}}}{N_{\mathrm{p}}}.$$

This is known as the **transformer equation**, and it simply states that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of loops in their coils.

The output voltage of a transformer can be less than, greater than, or equal to the input voltage, depending on the ratio of the number of loops in their coils. Some transformers even provide a variable output by allowing connection to be made at different points on the secondary coil. A **step-up**

transformer is one that increases voltage, whereas a **step-down transformer** decreases voltage. Assuming, as we have, that resistance is negligible, the electrical power output of a transformer equals its input. This is nearly true in practice—transformer efficiency often exceeds 99%. Equating the power input and output,

Equation:

$$P_{\mathrm{p}} = I_{\mathrm{p}} V_{\mathrm{p}} = I_{\mathrm{s}} V_{\mathrm{s}} = P_{\mathrm{s}}.$$

Rearranging terms gives

Equation:

$$\frac{V_{\mathrm{s}}}{V_{\mathrm{p}}} = \frac{I_{\mathrm{p}}}{I_{\mathrm{s}}}.$$

Combining this with $\frac{V_{\mathrm{s}}}{V_{\mathrm{p}}} = \frac{N_{\mathrm{s}}}{N_{\mathrm{p}}}$, we find that

Equation:

$$\frac{I_{\mathrm{s}}}{I_{\mathrm{p}}} = \frac{N_{\mathrm{p}}}{N_{\mathrm{s}}}$$

is the relationship between the output and input currents of a transformer. So if voltage increases, current decreases. Conversely, if voltage decreases, current increases.

Example:

Calculating Characteristics of a Step-Up Transformer

A portable x-ray unit has a step-up transformer, the 120 V input of which is transformed to the 100 kV output needed by the x-ray tube. The primary has 50 loops and draws a current of 10.00 A when in use. (a) What is the number of loops in the secondary? (b) Find the current output of the secondary.

Strategy and Solution for (a)

We solve $\frac{V_\mathrm{s}}{V_\mathrm{p}} = \frac{N_\mathrm{s}}{N_\mathrm{p}}$ for N_s, the number of loops in the secondary, and enter the known values. This gives

Equation:

$$\begin{aligned} N_\mathrm{s} &= N_\mathrm{p}\frac{V_\mathrm{s}}{V_\mathrm{p}} \\ &= (50)\frac{100{,}000\text{ V}}{120\text{ V}} = 4.17 \times 10^4. \end{aligned}$$

Discussion for (a)

A large number of loops in the secondary (compared with the primary) is required to produce such a large voltage. This would be true for neon sign transformers and those supplying high voltage inside TVs and CRTs.

Strategy and Solution for (b)

We can similarly find the output current of the secondary by solving $\frac{I_\mathrm{s}}{I_\mathrm{p}} = \frac{N_\mathrm{p}}{N_\mathrm{s}}$ for I_s and entering known values. This gives

Equation:

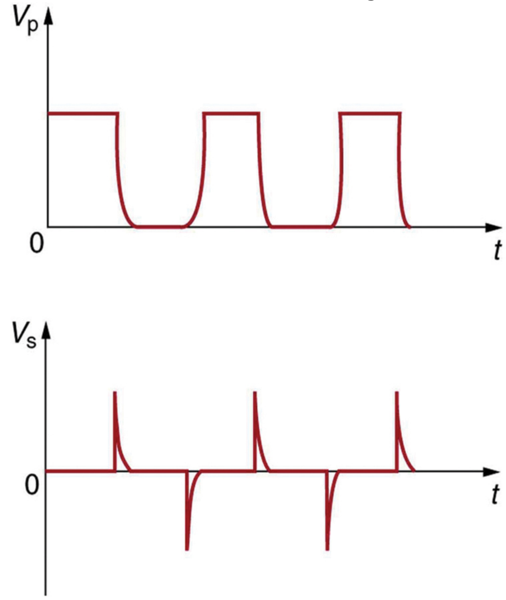
$$\begin{aligned} I_\mathrm{s} &= I_\mathrm{p}\frac{N_\mathrm{p}}{N_\mathrm{s}} \\ &= (10.00\text{ A})\,\frac{50}{4.17\times 10^4} = 12.0\text{ mA}. \end{aligned}$$

Discussion for (b)

As expected, the current output is significantly less than the input. In certain spectacular demonstrations, very large voltages are used to produce long arcs, but they are relatively safe because the transformer output does not supply a large current. Note that the power input here is $P_\mathrm{p} = I_\mathrm{p}V_\mathrm{p} = (10.00\text{ A})(120\text{ V}) = 1.20\text{ kW}$. This equals the power output $P_\mathrm{p} = I_\mathrm{s}V_\mathrm{s} = (12.0\text{ mA})(100\text{ kV}) = 1.20\text{ kW}$, as we assumed in the derivation of the equations used.

The fact that transformers are based on Faraday’s law of induction makes it clear why we cannot use transformers to change DC voltages. If there is no change in primary voltage, there is no voltage induced in the secondary. One possibility is to connect DC to the primary coil through a switch. As the switch is opened and closed, the secondary produces a voltage like that

in [link]. This is not really a practical alternative, and AC is in common use wherever it is necessary to increase or decrease voltages.



Transformers do not work for pure DC voltage input, but if it is switched on and off as on the top graph, the output will look something like that on the bottom graph. This is not the sinusoidal AC most AC appliances need.

Example:
Calculating Characteristics of a Step-Down Transformer
A battery charger meant for a series connection of ten nickel-cadmium batteries (total emf of 12.5 V DC) needs to have a 15.0 V output to charge the batteries. It uses a step-down transformer with a 200-loop primary and a 120 V input. (a) How many loops should there be in the secondary coil? (b) If the charging current is 16.0 A, what is the input current?

Strategy and Solution for (a)

You would expect the secondary to have a small number of loops. Solving $\frac{V_{\mathrm{s}}}{V_{\mathrm{p}}} = \frac{N_{\mathrm{s}}}{N_{\mathrm{p}}}$ for N_{s} and entering known values gives

Equation:

$$\begin{aligned} N_{\mathrm{s}} &= N_{\mathrm{p}} \frac{V_{\mathrm{s}}}{V_{\mathrm{p}}} \\ &= (200) \frac{15.0\ \mathrm{V}}{120\ \mathrm{V}} = 25. \end{aligned}$$

Strategy and Solution for (b)

The current input can be obtained by solving $\frac{I_{\mathrm{s}}}{I_{\mathrm{p}}} = \frac{N_{\mathrm{p}}}{N_{\mathrm{s}}}$ for I_{p} and entering known values. This gives

Equation:

$$\begin{aligned} I_{\mathrm{p}} &= I_{\mathrm{s}} \frac{N_{\mathrm{s}}}{N_{\mathrm{p}}} \\ &= (16.0\ \mathrm{A}) \frac{25}{200} = 2.00\ \mathrm{A}. \end{aligned}$$

Discussion

The number of loops in the secondary is small, as expected for a step-down transformer. We also see that a small input current produces a larger output current in a step-down transformer. When transformers are used to operate large magnets, they sometimes have a small number of very heavy loops in the secondary. This allows the secondary to have low internal resistance and produce large currents. Note again that this solution is based on the assumption of 100% efficiency—or power out equals power in ($P_{\mathrm{p}} = P_{\mathrm{s}}$)—reasonable for good transformers. In this case the primary and secondary power is 240 W. (Verify this for yourself as a consistency check.) Note that the Ni-Cd batteries need to be charged from a DC power source (as would a 12 V battery). So the AC output of the secondary coil needs to be converted into DC. This is done using something called a rectifier, which uses devices called diodes that allow only a one-way flow of current.

Transformers have many applications in electrical safety systems, which are discussed in Electrical Safety: Systems and Devices.

Note:
PhET Explorations: Generator
Generate electricity with a bar magnet! Discover the physics behind the phenomena by exploring magnets and how you can use them to make a bulb light.
https://archive.cnx.org/specials/1e9b7292-ae74-11e5-a9dc-c7c8521ba8e6/generator/#sim-generator

Section Summary

- Transformers use induction to transform voltages from one value to another.
- For a transformer, the voltages across the primary and secondary coils are related by
  Equation:
  $$\frac{V_\text{s}}{V_\text{p}} = \frac{N_\text{s}}{N_\text{p}},$$
  where V_p and V_s are the voltages across primary and secondary coils having N_p and N_s turns.
- The currents I_p and I_s in the primary and secondary coils are related by $\frac{I_\text{s}}{I_\text{p}} = \frac{N_\text{p}}{N_\text{s}}$.
- A step-up transformer increases voltage and decreases current, whereas a step-down transformer decreases voltage and increases current.

Conceptual Questions

Exercise:

Problem:

Explain what causes physical vibrations in transformers at twice the frequency of the AC power involved.

Problems & Exercises

Exercise:

Problem:

A plug-in transformer, like that in [link], supplies 9.00 V to a video game system. (a) How many turns are in its secondary coil, if its input voltage is 120 V and the primary coil has 400 turns? (b) What is its input current when its output is 1.30 A?

Solution:

(a) 30.0

(b) 9.75×10^{-2} A

Exercise:

Problem:

An American traveler in New Zealand carries a transformer to convert New Zealand's standard 240 V to 120 V so that she can use some small appliances on her trip. (a) What is the ratio of turns in the primary and secondary coils of her transformer? (b) What is the ratio of input to output current? (c) How could a New Zealander traveling in the United States use this same transformer to power her 240 V appliances from 120 V?

Exercise:

Problem:

A cassette recorder uses a plug-in transformer to convert 120 V to 12.0 V, with a maximum current output of 200 mA. (a) What is the current input? (b) What is the power input? (c) Is this amount of power reasonable for a small appliance?

Solution:

(a) 20.0 mA

(b) 2.40 W

(c) Yes, this amount of power is quite reasonable for a small appliance.

Exercise:

Problem:

(a) What is the voltage output of a transformer used for rechargeable flashlight batteries, if its primary has 500 turns, its secondary 4 turns, and the input voltage is 120 V? (b) What input current is required to produce a 4.00 A output? (c) What is the power input?

Exercise:

Problem:

(a) The plug-in transformer for a laptop computer puts out 7.50 V and can supply a maximum current of 2.00 A. What is the maximum input current if the input voltage is 240 V? Assume 100% efficiency. (b) If the actual efficiency is less than 100%, would the input current need to be greater or smaller? Explain.

Solution:

(a) 0.063 A

(b) Greater input current needed.

Exercise:

Problem:

A multipurpose transformer has a secondary coil with several points at which a voltage can be extracted, giving outputs of 5.60, 12.0, and 480 V. (a) The input voltage is 240 V to a primary coil of 280 turns. What are the numbers of turns in the parts of the secondary used to produce the output voltages? (b) If the maximum input current is 5.00 A, what are the maximum output currents (each used alone)?

Exercise:

Problem:

A large power plant generates electricity at 12.0 kV. Its old transformer once converted the voltage to 335 kV. The secondary of this transformer is being replaced so that its output can be 750 kV for more efficient cross-country transmission on upgraded transmission lines. (a) What is the ratio of turns in the new secondary compared with the old secondary? (b) What is the ratio of new current output to old output (at 335 kV) for the same power? (c) If the upgraded transmission lines have the same resistance, what is the ratio of new line power loss to old?

Solution:

(a) 2.2

(b) 0.45

(c) 0.20, or 20.0%

Exercise:

Problem:

If the power output in the previous problem is 1000 MW and line resistance is $2.00\ \Omega$, what were the old and new line losses?

Exercise:

Problem: Unreasonable Results

The 335 kV AC electricity from a power transmission line is fed into the primary coil of a transformer. The ratio of the number of turns in the secondary to the number in the primary is $N_{\text{s}}/N_{\text{p}} = 1000$. (a) What voltage is induced in the secondary? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Solution:

(a) 335 MV

(b) way too high, well beyond the breakdown voltage of air over reasonable distances

(c) input voltage is too high

Exercise:

Problem: Construct Your Own Problem

Consider a double transformer to be used to create very large voltages. The device consists of two stages. The first is a transformer that produces a much larger output voltage than its input. The output of the first transformer is used as input to a second transformer that further increases the voltage. Construct a problem in which you calculate the output voltage of the final stage based on the input voltage of the first stage and the number of turns or loops in both parts of both transformers (four coils in all). Also calculate the maximum output current of the final stage based on the input current. Discuss the possibility of power losses in the devices and the effect on the output current and power.

Glossary

transformer
: a device that transforms voltages from one value to another using induction

transformer equation
: the equation showing that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of loops in their coils; $\frac{V_\text{s}}{V_\text{p}} = \frac{N_\text{s}}{N_\text{p}}$

step-up transformer
: a transformer that increases voltage

step-down transformer

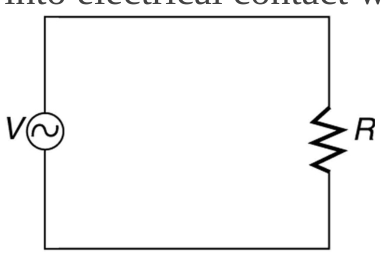
a transformer that decreases voltage

Electrical Safety: Systems and Devices

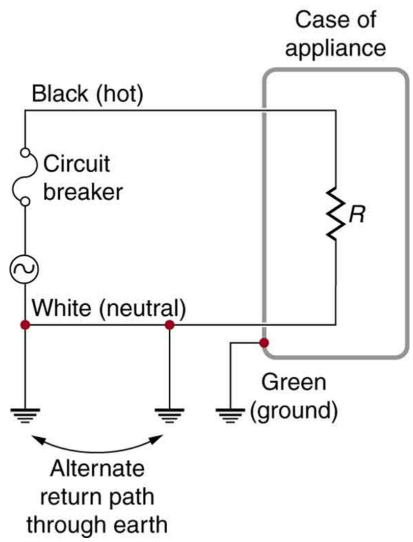
- Explain how various modern safety features in electric circuits work, with an emphasis on how induction is employed.

Electricity has two hazards. A **thermal hazard** occurs when there is electrical overheating. A **shock hazard** occurs when electric current passes through a person. Both hazards have already been discussed. Here we will concentrate on systems and devices that prevent electrical hazards.

[link] shows the schematic for a simple AC circuit with no safety features. This is not how power is distributed in practice. Modern household and industrial wiring requires the **three-wire system**, shown schematically in [link], which has several safety features. First is the familiar *circuit breaker* (or *fuse*) to prevent thermal overload. Second, there is a protective *case* around the appliance, such as a toaster or refrigerator. The case's safety feature is that it prevents a person from touching exposed wires and coming into electrical contact with the circuit, helping prevent shocks.



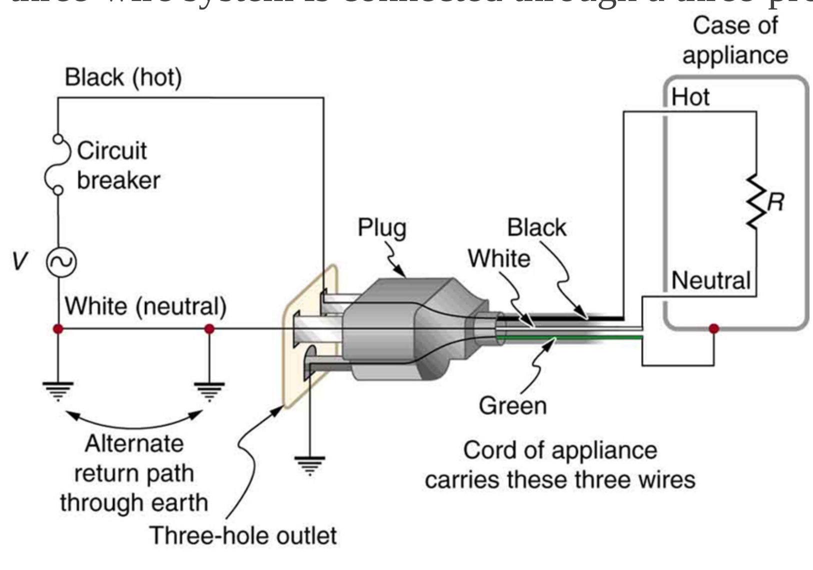
Schematic of a simple AC circuit with a voltage source and a single appliance represented by the resistance R. There are no safety features in this circuit.



The three-wire system connects the neutral wire to the earth at the voltage source and user location, forcing it to be at zero volts and supplying an alternative return path for the current through the earth. Also grounded to zero volts is the case of the appliance. A circuit breaker or fuse protects against thermal overload and is in series on the active (live/hot) wire. Note that wire insulation colors vary with region and it is essential to

check locally to determine which color codes are in use (and even if they were followed in the particular installation).

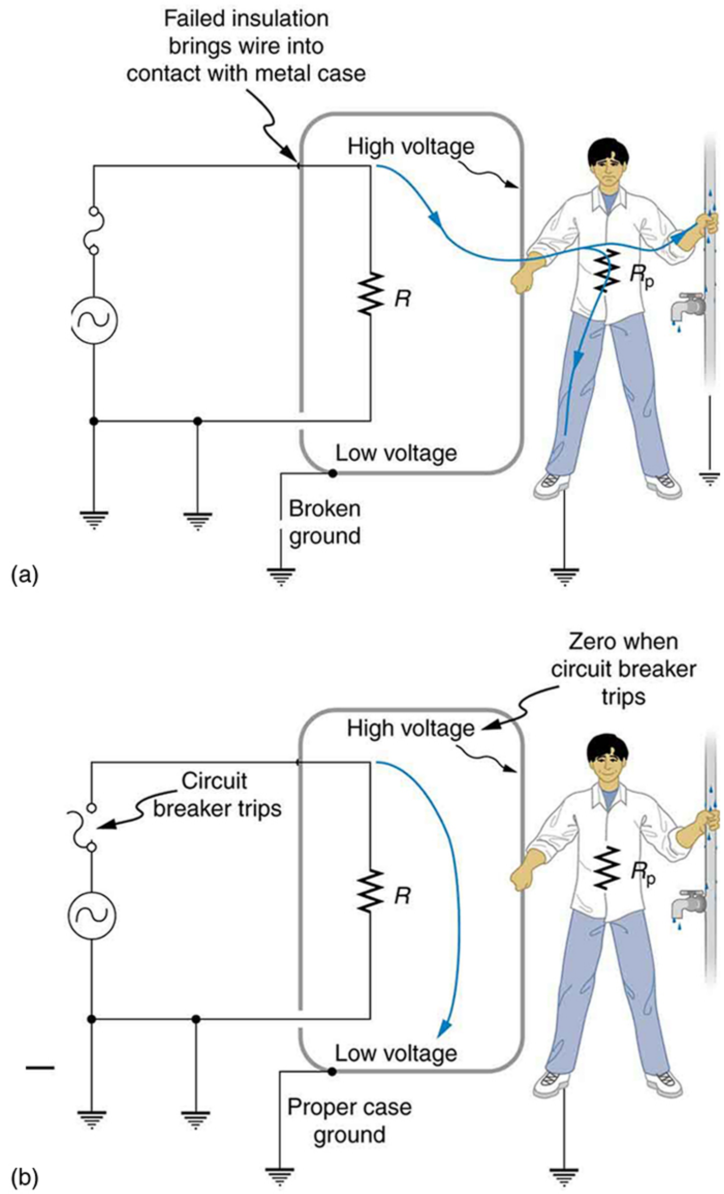
There are *three connections to earth or ground* (hereafter referred to as "earth/ground") shown in [link]. Recall that an earth/ground connection is a low-resistance path directly to the earth. The two earth/ground connections on the *neutral wire* force it to be at zero volts relative to the earth, giving the wire its name. This wire is therefore safe to touch even if its insulation, usually white, is missing. The neutral wire is the return path for the current to follow to complete the circuit. Furthermore, the two earth/ground connections supply an alternative path through the earth, a good conductor, to complete the circuit. The earth/ground connection closest to the power source could be at the generating plant, while the other is at the user's location. The third earth/ground is to the case of the appliance, through the green *earth/ground wire*, forcing the case, too, to be at zero volts. The *live* or *hot wire* (hereafter referred to as "live/hot") supplies voltage and current to operate the appliance. [link] shows a more pictorial version of how the three-wire system is connected through a three-prong plug to an appliance.



The standard three-prong plug can only be inserted in one way, to assure proper function of the three-wire system.

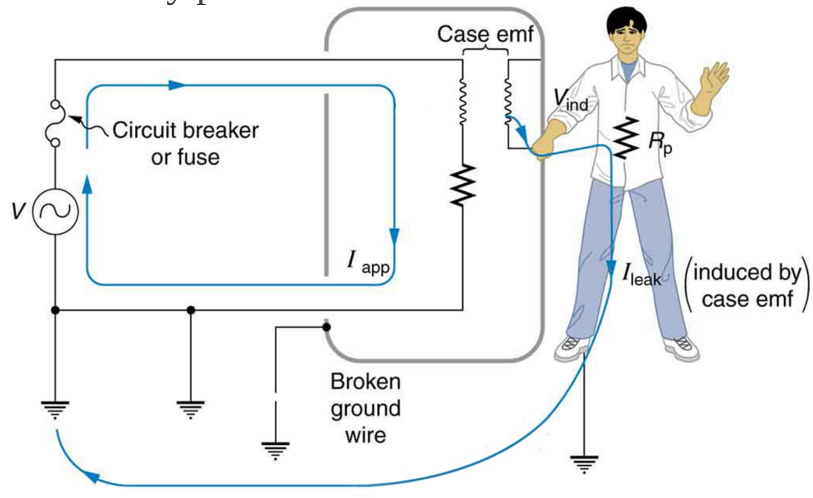
A note on insulation color-coding: Insulating plastic is color-coded to identify live/hot, neutral and ground wires but these codes vary around the world. Live/hot wires may be brown, red, black, blue or grey. Neutral wire may be blue, black or white. Since the same color may be used for live/hot or neutral in different parts of the world, it is essential to determine the color code in your region. The only exception is the earth/ground wire which is often green but may be yellow or just bare wire. Striped coatings are sometimes used for the benefit of those who are colorblind.

The three-wire system replaced the older two-wire system, which lacks an earth/ground wire. Under ordinary circumstances, insulation on the live/hot and neutral wires prevents the case from being directly in the circuit, so that the earth/ground wire may seem like double protection. Grounding the case solves more than one problem, however. The simplest problem is worn insulation on the live/hot wire that allows it to contact the case, as shown in [link]. Lacking an earth/ground connection (some people cut the third prong off the plug because they only have outdated two hole receptacles), a severe shock is possible. This is particularly dangerous in the kitchen, where a good connection to earth/ground is available through water on the floor or a water faucet. With the earth/ground connection intact, the circuit breaker will trip, forcing repair of the appliance. Why are some appliances still sold with two-prong plugs? These have nonconducting cases, such as power tools with impact resistant plastic cases, and are called *doubly insulated*. Modern two-prong plugs can be inserted into the asymmetric standard outlet in only one way, to ensure proper connection of live/hot and neutral wires.



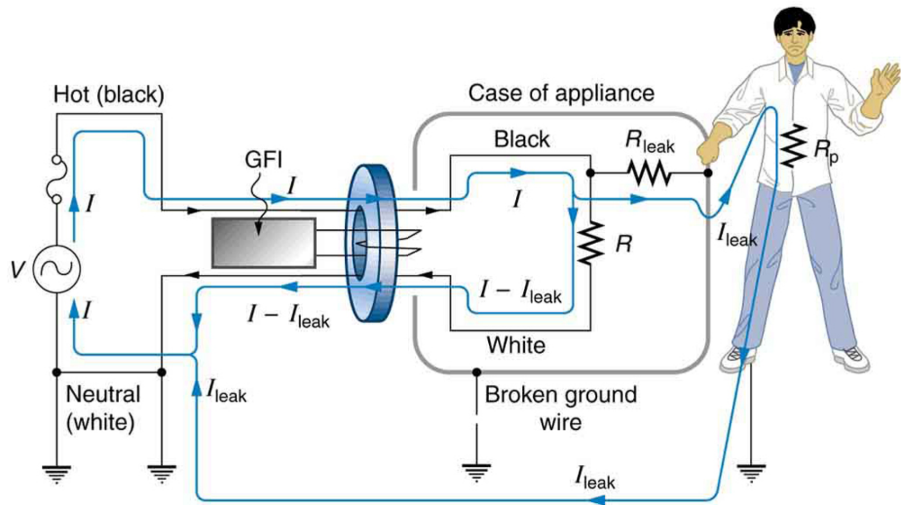
Worn insulation allows the live/hot wire to come into direct contact with the metal case of this appliance. (a) The earth/ground connection being broken, the person is severely shocked. The appliance may operate normally in this situation. (b) With a proper earth/ground, the circuit breaker trips, forcing repair of the appliance.

Electromagnetic induction causes a more subtle problem that is solved by grounding the case. The AC current in appliances can induce an emf on the case. If grounded, the case voltage is kept near zero, but if the case is not grounded, a shock can occur as pictured in [link]. Current driven by the induced case emf is called a *leakage current*, although current does not necessarily pass from the resistor to the case.



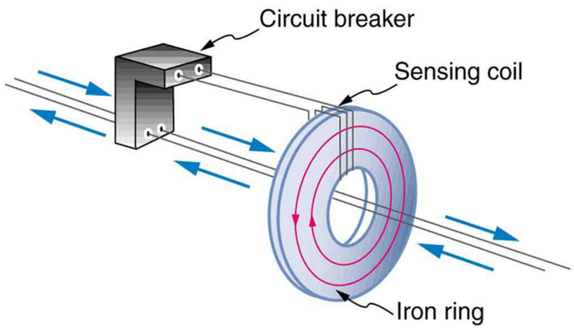
AC currents can induce an emf on the case of an appliance. The voltage can be large enough to cause a shock. If the case is grounded, the induced emf is kept near zero.

A *ground fault interrupter* (GFI) is a safety device found in updated kitchen and bathroom wiring that works based on electromagnetic induction. GFIs compare the currents in the live/hot and neutral wires. When live/hot and neutral currents are not equal, it is almost always because current in the neutral is less than in the live/hot wire. Then some of the current, again called a leakage current, is returning to the voltage source by a path other than through the neutral wire. It is assumed that this path presents a hazard, such as shown in [link]. GFIs are usually set to interrupt the circuit if the leakage current is greater than 5 mA, the accepted maximum harmless shock. Even if the leakage current goes safely to earth/ground through an intact earth/ground wire, the GFI will trip, forcing repair of the leakage.



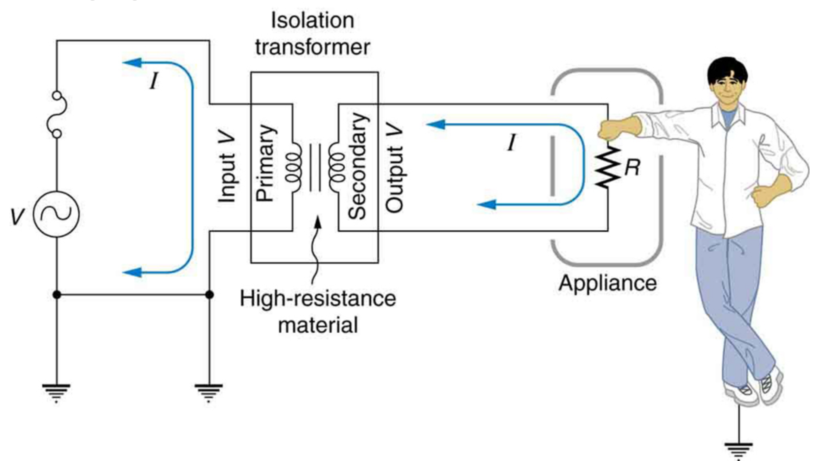
A ground fault interrupter (GFI) compares the currents in the live/hot and neutral wires and will trip if their difference exceeds a safe value. The leakage current here follows a hazardous path that could have been prevented by an intact earth/ground wire.

[link] shows how a GFI works. If the currents in the live/hot and neutral wires are equal, then they induce equal and opposite emfs in the coil. If not, then the circuit breaker will trip.



A GFI compares currents by using both to induce an emf in the same coil. If the currents are equal, they will induce equal but opposite emfs.

Another induction-based safety device is the *isolation transformer*, shown in [link]. Most isolation transformers have equal input and output voltages. Their function is to put a large resistance between the original voltage source and the device being operated. This prevents a complete circuit between them, even in the circumstance shown. There is a complete circuit through the appliance. But there is not a complete circuit for current to flow through the person in the figure, who is touching only one of the transformer's output wires, and neither output wire is grounded. The appliance is isolated from the original voltage source by the high resistance of the material between the transformer coils, hence the name isolation transformer. For current to flow through the person, it must pass through the high-resistance material between the coils, through the wire, the person, and back through the earth—a path with such a large resistance that the current is negligible.



An isolation transformer puts a large resistance between the original voltage source and the device, preventing a complete circuit between them.

The basics of electrical safety presented here help prevent many electrical hazards. Electrical safety can be pursued to greater depths. There are, for

example, problems related to different earth/ground connections for appliances in close proximity. Many other examples are found in hospitals. Microshock-sensitive patients, for instance, require special protection. For these people, currents as low as 0.1 mA may cause ventricular fibrillation. The interested reader can use the material presented here as a basis for further study.

Section Summary

- Electrical safety systems and devices are employed to prevent thermal and shock hazards.
- Circuit breakers and fuses interrupt excessive currents to prevent thermal hazards.
- The three-wire system guards against thermal and shock hazards, utilizing live/hot, neutral, and earth/ground wires, and grounding the neutral wire and case of the appliance.
- A ground fault interrupter (GFI) prevents shock by detecting the loss of current to unintentional paths.
- An isolation transformer insulates the device being powered from the original source, also to prevent shock.
- Many of these devices use induction to perform their basic function.

Conceptual Questions

Exercise:

Problem:

Does plastic insulation on live/hot wires prevent shock hazards, thermal hazards, or both?

Exercise:

Problem:

Why are ordinary circuit breakers and fuses ineffective in preventing shocks?

Exercise:

Problem:

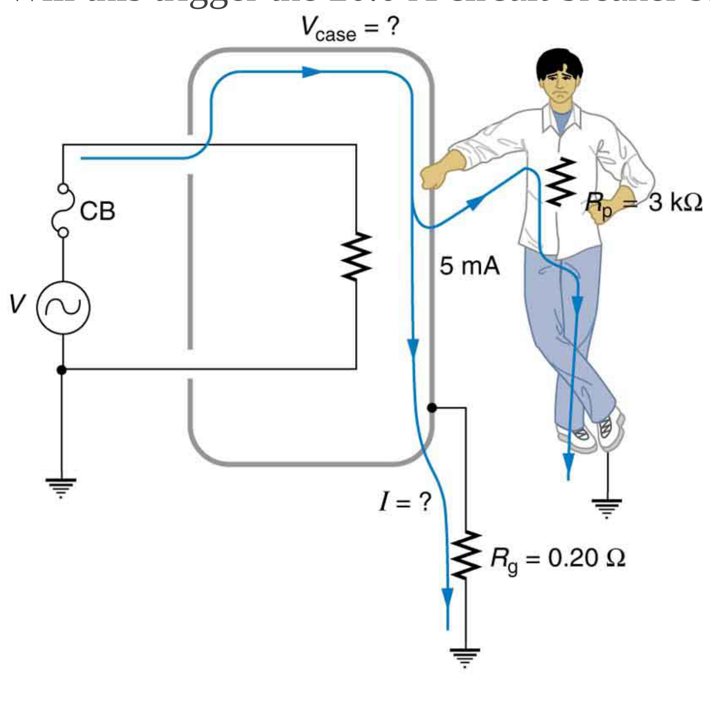
A GFI may trip just because the live/hot and neutral wires connected to it are significantly different in length. Explain why.

Problems & Exercises

Exercise:

Problem: **Integrated Concepts**

A short circuit to the grounded metal case of an appliance occurs as shown in [link]. The person touching the case is wet and only has a $3.00\text{ k}\Omega$ resistance to earth/ground. (a) What is the voltage on the case if 5.00 mA flows through the person? (b) What is the current in the short circuit if the resistance of the earth/ground wire is $0.200\ \Omega$? (c) Will this trigger the 20.0 A circuit breaker supplying the appliance?



A person can be shocked even when the case of an appliance is grounded. The large short circuit current

produces a voltage on the case of the appliance, since the resistance of the earth/ground wire is not zero.

Solution:

(a) 15.0 V

(b) 75.0 A

(c) yes

Glossary

thermal hazard
: the term for electrical hazards due to overheating

shock hazard
: the term for electrical hazards due to current passing through a human

three-wire system
: the wiring system used at present for safety reasons, with live, neutral, and ground wires

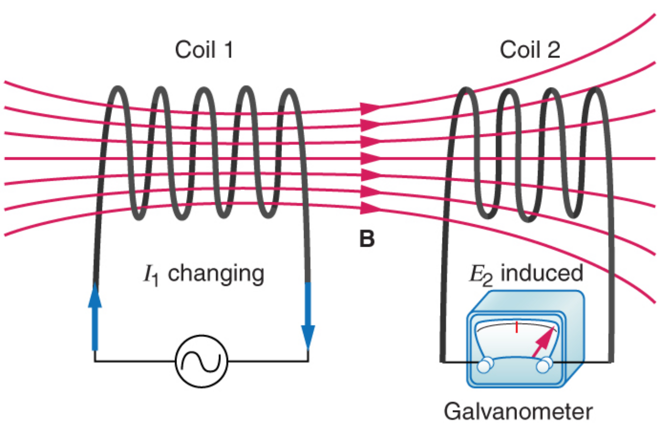
Inductance

- Calculate the inductance of an inductor.
- Calculate the energy stored in an inductor.
- Calculate the emf generated in an inductor.

Inductors

Induction is the process in which an emf is induced by changing magnetic flux. Many examples have been discussed so far, some more effective than others. Transformers, for example, are designed to be particularly effective at inducing a desired voltage and current with very little loss of energy to other forms. Is there a useful physical quantity related to how "effective" a given device is? The answer is yes, and that physical quantity is called **inductance**.

Mutual inductance is the effect of Faraday's law of induction for one device upon another, such as the primary coil in transmitting energy to the secondary in a transformer. See [link], where simple coils induce emfs in one another.



These coils can induce emfs in one another like an inefficient transformer. Their mutual inductance M indicates the

effectiveness of the coupling between them. Here a change in current in coil 1 is seen to induce an emf in coil 2. (Note that "E_2 induced" represents the induced emf in coil 2.)

In the many cases where the geometry of the devices is fixed, flux is changed by varying current. We therefore concentrate on the rate of change of current, $\Delta I/\Delta t$, as the cause of induction. A change in the current I_1 in one device, coil 1 in the figure, induces an emf_2 in the other. We express this in equation form as

Equation:

$$\text{emf}_2 = -M\frac{\Delta I_1}{\Delta t},$$

where M is defined to be the mutual inductance between the two devices. The minus sign is an expression of Lenz's law. The larger the mutual inductance M, the more effective the coupling. For example, the coils in [link] have a small M compared with the transformer coils in [link]. Units for M are $(\text{V} \cdot \text{s})/\text{A} = \Omega \cdot \text{s}$, which is named a **henry** (H), after Joseph Henry. That is, $1 \text{ H} = 1 \ \Omega \cdot \text{s}$.

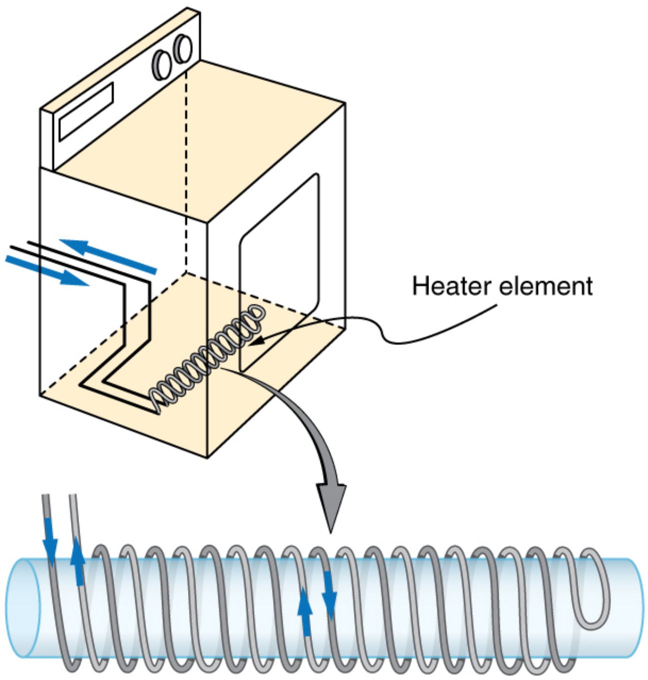
Nature is symmetric here. If we change the current I_2 in coil 2, we induce an emf_1 in coil 1, which is given by

Equation:

$$\text{emf}_1 = -M\frac{\Delta I_2}{\Delta t},$$

where M is the same as for the reverse process. Transformers run backward with the same effectiveness, or mutual inductance M.

A large mutual inductance M may or may not be desirable. We want a transformer to have a large mutual inductance. But an appliance, such as an electric clothes dryer, can induce a dangerous emf on its case if the mutual inductance between its coils and the case is large. One way to reduce mutual inductance M is to counterwind coils to cancel the magnetic field produced. (See [link].)



The heating coils of an electric clothes dryer can be counter-wound so that their magnetic fields cancel one another, greatly reducing the mutual inductance with the case of the dryer.

Self-inductance, the effect of Faraday’s law of induction of a device on itself, also exists. When, for example, current through a coil is increased, the magnetic field and flux also increase, inducing a counter emf, as

required by Lenz's law. Conversely, if the current is decreased, an emf is induced that opposes the decrease. Most devices have a fixed geometry, and so the change in flux is due entirely to the change in current ΔI through the device. The induced emf is related to the physical geometry of the device and the rate of change of current. It is given by

Equation:

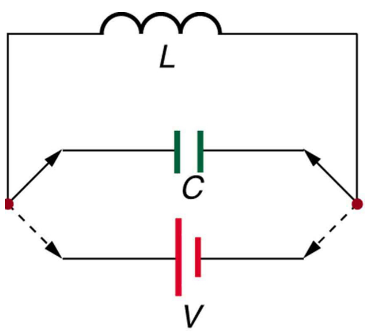
$$\text{emf} = -L\frac{\Delta I}{\Delta t},$$

where L is the self-inductance of the device. A device that exhibits significant self-inductance is called an **inductor**, and given the symbol in [link].

The minus sign is an expression of Lenz's law, indicating that emf opposes the change in current. Units of self-inductance are henries (H) just as for mutual inductance. The larger the self-inductance L of a device, the greater its opposition to any change in current through it. For example, a large coil with many turns and an iron core has a large L and will not allow current to change quickly. To avoid this effect, a small L must be achieved, such as by counterwinding coils as in [link].

A 1 H inductor is a large inductor. To illustrate this, consider a device with $L = 1.0\text{ H}$ that has a 10 A current flowing through it. What happens if we try to shut off the current rapidly, perhaps in only 1.0 ms? An emf, given by $\text{emf} = -L(\Delta I/\Delta t)$, will oppose the change. Thus an emf will be induced given by $\text{emf} = -L(\Delta I/\Delta t) = (1.0\text{ H})[(10\text{ A})/(1.0\text{ ms})] = 10{,}000\text{ V}$. The positive sign means this large voltage is in the same direction as the current, opposing its decrease. Such large emfs can cause arcs, damaging switching equipment, and so it may be necessary to change current more slowly.

There are uses for such a large induced voltage. Camera flashes use a battery, two inductors that function as a transformer, and a switching system or oscillator to induce large voltages. (Remember that we need a changing magnetic field, brought about by a changing current, to induce a voltage in another coil.) The oscillator system will do this many times as the battery voltage is boosted to over one thousand volts. (You may hear the high pitched whine from the transformer as the capacitor is being charged.) A capacitor stores the high voltage for later use in powering the flash. (See [link].)



Through rapid switching of an inductor, 1.5 V batteries can be used to induce emfs of several thousand volts. This voltage can be used to store charge in a capacitor for later use, such as in a camera flash attachment.

It is possible to calculate L for an inductor given its geometry (size and shape) and knowing the magnetic field that it produces. This is difficult in most cases, because of the complexity of the field created. So in this text the inductance L is usually a given quantity. One exception is the solenoid, because it has a very uniform field inside, a nearly zero field outside, and a simple shape. It is instructive to derive an equation for its inductance. We start by noting that the induced emf is given by Faraday's law of induction as $\text{emf} = -N(\Delta\Phi/\Delta t)$ and, by the definition of self-inductance, as $\text{emf} = -L(\Delta I/\Delta t)$. Equating these yields

Equation:

$$\text{emf} = -N\frac{\Delta\Phi}{\Delta t} = -L\frac{\Delta I}{\Delta t}.$$

Solving for L gives

Equation:

$$L = N\frac{\Delta\Phi}{\Delta I}.$$

This equation for the self-inductance L of a device is always valid. It means that self-inductance L depends on how effective the current is in creating flux; the more effective, the greater $\Delta\Phi$/ ΔI is.

Let us use this last equation to find an expression for the inductance of a solenoid. Since the area A of a solenoid is fixed, the change in flux is $\Delta\Phi = \Delta(BA) = A\Delta B$. To find ΔB, we note that the magnetic field of a solenoid is given by $B = \mu_0\text{nI} = \mu_0\frac{\text{NI}}{\ell}$. (Here $n = N/\ell$, where N is the number of coils and ℓ is the solenoid's length.) Only the current changes, so that $\Delta\Phi = A\Delta B = \mu_0\text{NA}\frac{\Delta I}{\ell}$. Substituting $\Delta\Phi$ into $L = N\frac{\Delta\Phi}{\Delta I}$ gives

Equation:

$$L = N\frac{\Delta\Phi}{\Delta I} = N\frac{\mu_0\text{NA}\frac{\Delta I}{\ell}}{\Delta I}.$$

This simplifies to
Equation:

$$L = \frac{\mu_0 N^2 A}{\ell} \text{(solenoid)}.$$

This is the self-inductance of a solenoid of cross-sectional area A and length ℓ. Note that the inductance depends only on the physical characteristics of the solenoid, consistent with its definition.

Example:
Calculating the Self-inductance of a Moderate Size Solenoid
Calculate the self-inductance of a 10.0 cm long, 4.00 cm diameter solenoid that has 200 coils.
Strategy
This is a straightforward application of $L = \frac{\mu_0 N^2 A}{\ell}$, since all quantities in the equation except L are known.
Solution
Use the following expression for the self-inductance of a solenoid:
Equation:

$$L = \frac{\mu_0 N^2 A}{\ell}.$$

The cross-sectional area in this example is $A = \pi r^2 = (3.14...)(0.0200 \text{ m})^2 = 1.26 \times 10^{-3} \text{ m}^2$, N is given to be 200, and the length ℓ is 0.100 m. We know the permeability of free space is $\mu_0 = 4\pi \times 10^{-7} \text{ T} \cdot \text{m/A}$. Substituting these into the expression for L gives
Equation:

$$\begin{aligned} L &= \frac{(4\pi \times 10^{-7} \text{ T} \cdot \text{m/A})(200)^2 (1.26 \times 10^{-3} \text{ m}^2)}{0.100 \text{ m}} \\ &= 0.632 \text{ mH}. \end{aligned}$$

Discussion

This solenoid is moderate in size. Its inductance of nearly a millihenry is also considered moderate.

One common application of inductance is used in traffic lights that can tell when vehicles are waiting at the intersection. An electrical circuit with an inductor is placed in the road under the place a waiting car will stop over. The body of the car increases the inductance and the circuit changes sending a signal to the traffic lights to change colors. Similarly, metal detectors used for airport security employ the same technique. A coil or inductor in the metal detector frame acts as both a transmitter and a receiver. The pulsed signal in the transmitter coil induces a signal in the receiver. The self-inductance of the circuit is affected by any metal object in the path. Such detectors can be adjusted for sensitivity and also can indicate the approximate location of metal found on a person. (But they will not be able to detect any plastic explosive such as that found on the “underwear bomber.”) See [link].

The familiar security gate at an airport can not only detect metals but also indicate their approximate height above the floor.

(credit: Alexbuirds, Wikimedia Commons)

Energy Stored in an Inductor

We know from Lenz's law that inductances oppose changes in current. There is an alternative way to look at this opposition that is based on energy. Energy is stored in a magnetic field. It takes time to build up energy, and it also takes time to deplete energy; hence, there is an opposition to rapid change. In an inductor, the magnetic field is directly proportional to current and to the inductance of the device. It can be shown that the **energy stored in an inductor** E_{ind} is given by

Equation:

$$E_{\text{ind}} = \frac{1}{2} L I^2.$$

This expression is similar to that for the energy stored in a capacitor.

Example:

Calculating the Energy Stored in the Field of a Solenoid

How much energy is stored in the 0.632 mH inductor of the preceding example when a 30.0 A current flows through it?

Strategy

The energy is given by the equation $E_{\text{ind}} = \frac{1}{2} L I^2$, and all quantities except E_{ind} are known.

Solution

Substituting the value for L found in the previous example and the given current into $E_{\text{ind}} = \frac{1}{2} L I^2$ gives

Equation:

$$\begin{aligned} E_{\text{ind}} &= \tfrac{1}{2} L I^2 \\ &= 0.5(0.632 \times 10^{-3}\ \text{H})(30.0\ \text{A})^2 = 0.284\ \text{J}. \end{aligned}$$

Discussion
This amount of energy is certainly enough to cause a spark if the current is suddenly switched off. It cannot be built up instantaneously unless the power input is infinite.

Section Summary

- Inductance is the property of a device that tells how effectively it induces an emf in another device.
- Mutual inductance is the effect of two devices in inducing emfs in each other.
- A change in current $\Delta I_1/\Delta t$ in one induces an emf emf_2 in the second:
  Equation:
  $$\text{emf}_2 = -M\frac{\Delta I_1}{\Delta t},$$
  where M is defined to be the mutual inductance between the two devices, and the minus sign is due to Lenz's law.
- Symmetrically, a change in current $\Delta I_2/\Delta t$ through the second device induces an emf emf_1 in the first:
  Equation:
  $$\text{emf}_1 = -M\frac{\Delta I_2}{\Delta t},$$
  where M is the same mutual inductance as in the reverse process.
- Current changes in a device induce an emf in the device itself.
- Self-inductance is the effect of the device inducing emf in itself.
- The device is called an inductor, and the emf induced in it by a change in current through it is
  Equation:
  $$\text{emf} = -L\frac{\Delta I}{\Delta t},$$

where L is the self-inductance of the inductor, and $\Delta I/\Delta t$ is the rate of change of current through it. The minus sign indicates that emf opposes the change in current, as required by Lenz's law.

- The unit of self- and mutual inductance is the henry (H), where $1\ \text{H} = 1\ \Omega \cdot \text{s}$.
- The self-inductance L of an inductor is proportional to how much flux changes with current. For an N-turn inductor,
  Equation:

$$L = N\frac{\Delta\Phi}{\Delta I}.$$

- The self-inductance of a solenoid is
  Equation:

$$L = \frac{\mu_0 N^2 A}{\ell}\,(\text{solenoid}),$$

  where N is its number of turns in the solenoid, A is its cross-sectional area, ℓ is its length, and $\mu_0 = 4\pi \times 10^{-7}\ \text{T}\cdot\text{m/A}$ is the permeability of free space.
- The energy stored in an inductor E_{ind} is
  Equation:

$$E_{\text{ind}} = \frac{1}{2}LI^2.$$

Conceptual Questions

Exercise:

Problem:

How would you place two identical flat coils in contact so that they had the greatest mutual inductance? The least?

Exercise:

Problem:

How would you shape a given length of wire to give it the greatest self-inductance? The least?

Exercise:

Problem:

Verify, as was concluded without proof in [link], that units of $\text{T} \cdot \text{m}^2/A = \Omega \cdot \text{s} = \text{H}$.

Problems & Exercises

Exercise:

Problem:

Two coils are placed close together in a physics lab to demonstrate Faraday's law of induction. A current of 5.00 A in one is switched off in 1.00 ms, inducing a 9.00 V emf in the other. What is their mutual inductance?

Solution:

1.80 mH

Exercise:

Problem:

If two coils placed next to one another have a mutual inductance of 5.00 mH, what voltage is induced in one when the 2.00 A current in the other is switched off in 30.0 ms?

Exercise:

Problem:

The 4.00 A current through a 7.50 mH inductor is switched off in 8.33 ms. What is the emf induced opposing this?

Solution:

3.60 V

Exercise:

Problem:

A device is turned on and 3.00 A flows through it 0.100 ms later. What is the self-inductance of the device if an induced 150 V emf opposes this?

Exercise:

Problem:

Starting with $\text{emf}_2 = -M\frac{\Delta I_1}{\Delta t}$, show that the units of inductance are $(\text{V}\cdot\text{s})/\text{A} = \Omega\cdot\text{s}$.

Exercise:

Problem:

Camera flashes charge a capacitor to high voltage by switching the current through an inductor on and off rapidly. In what time must the 0.100 A current through a 2.00 mH inductor be switched on or off to induce a 500 V emf?

Exercise:

Problem:

A large research solenoid has a self-inductance of 25.0 H. (a) What induced emf opposes shutting it off when 100 A of current through it is switched off in 80.0 ms? (b) How much energy is stored in the inductor at full current? (c) At what rate in watts must energy be dissipated to switch the current off in 80.0 ms? (d) In view of the answer to the last part, is it surprising that shutting it down this quickly is difficult?

Solution:

(a) 31.3 kV

(b) 125 kJ

(c) 1.56 MW

(d) No, it is not surprising since this power is very high.

Exercise:

Problem:

(a) Calculate the self-inductance of a 50.0 cm long, 10.0 cm diameter solenoid having 1000 loops. (b) How much energy is stored in this inductor when 20.0 A of current flows through it? (c) How fast can it be turned off if the induced emf cannot exceed 3.00 V?

Exercise:

Problem:

A precision laboratory resistor is made of a coil of wire 1.50 cm in diameter and 4.00 cm long, and it has 500 turns. (a) What is its self-inductance? (b) What average emf is induced if the 12.0 A current through it is turned on in 5.00 ms (one-fourth of a cycle for 50 Hz AC)? (c) What is its inductance if it is shortened to half its length and counter-wound (two layers of 250 turns in opposite directions)?

Solution:

(a) 1.39 mH

(b) 3.33 V

(c) Zero

Exercise:

Problem:

The heating coils in a hair dryer are 0.800 cm in diameter, have a combined length of 1.00 m, and a total of 400 turns. (a) What is their total self-inductance assuming they act like a single solenoid? (b) How much energy is stored in them when 6.00 A flows? (c) What average emf opposes shutting them off if this is done in 5.00 ms (one-fourth of a cycle for 50 Hz AC)?

Exercise:

Problem:

When the 20.0 A current through an inductor is turned off in 1.50 ms, an 800 V emf is induced, opposing the change. What is the value of the self-inductance?

Solution:

60.0 mH

Exercise:

Problem:

How fast can the 150 A current through a 0.250 H inductor be shut off if the induced emf cannot exceed 75.0 V?

Exercise:

Problem: Integrated Concepts

A very large, superconducting solenoid such as one used in MRI scans, stores 1.00 MJ of energy in its magnetic field when 100 A flows. (a) Find its self-inductance. (b) If the coils "go normal," they gain resistance and start to dissipate thermal energy. What temperature increase is produced if all the stored energy goes into heating the 1000 kg magnet, given its average specific heat is $200\ \text{J/kg}\cdot\text{ºC}$?

Solution:

(a) 200 H

(b) 5.00°C

Exercise:

Problem: **Unreasonable Results**

A 25.0 H inductor has 100 A of current turned off in 1.00 ms. (a) What voltage is induced to oppose this? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Glossary

inductance
: a property of a device describing how efficient it is at inducing emf in another device

mutual inductance
: how effective a pair of devices are at inducing emfs in each other

henry
: the unit of inductance; $1\text{ H}=1\ \Omega\cdot\text{s}$

self-inductance
: how effective a device is at inducing emf in itself

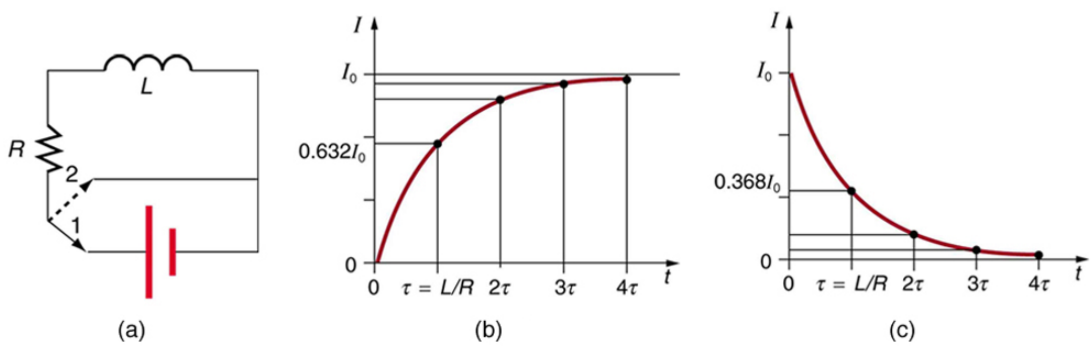
inductor
: a device that exhibits significant self-inductance

energy stored in an inductor
: self-explanatory; calculated by $E_{\text{ind}}=\frac{1}{2}LI^2$

RL Circuits

- Calculate the current in an RL circuit after a specified number of characteristic time steps.
- Calculate the characteristic time of an RL circuit.
- Sketch the current in an RL circuit over time.

We know that the current through an inductor L cannot be turned on or off instantaneously. The change in current changes flux, inducing an emf opposing the change (Lenz's law). How long does the opposition last? Current *will* flow and *can* be turned off, but how long does it take? [link] shows a switching circuit that can be used to examine current through an inductor as a function of time.



(a) An *RL* circuit with a switch to turn current on and off. When in position 1, the battery, resistor, and inductor are in series and a current is established. In position 2, the battery is removed and the current eventually stops because of energy loss in the resistor. (b) A graph of current growth versus time when the switch is moved to position 1. (c) A graph of current decay when the switch is moved to position 2.

When the switch is first moved to position 1 (at $t = 0$), the current is zero and it eventually rises to $I_0 = \mathrm{V/R}$, where R is the total resistance of the circuit. The opposition of the inductor L is greatest at the beginning, because the amount of change is greatest. The opposition it poses is in the form of an induced emf, which decreases to zero as the current approaches its final value. The opposing emf is proportional to the amount of change

left. This is the hallmark of an exponential behavior, and it can be shown with calculus that

Equation:

$$I = I_0(1 - e^{-t/\tau}) \quad \text{(turning on)},$$

is the current in an *RL* circuit when switched on (Note the similarity to the exponential behavior of the voltage on a charging capacitor). The initial current is zero and approaches $I_0 = \mathrm{V/R}$ with a **characteristic time constant** τ for an *RL* circuit, given by

Equation:

$$\tau = \frac{L}{R},$$

where τ has units of seconds, since $1\ \mathrm{H}{=}1\ \Omega{\cdot}\mathrm{s}$. In the first period of time τ, the current rises from zero to $0.632 I_0$, since $I = I_0(1 - e^{-1}) = I_0(1 - 0.368) = 0.632 I_0$. The current will go 0.632 of the remainder in the next time τ. A well-known property of the exponential is that the final value is never exactly reached, but 0.632 of the remainder to that value is achieved in every characteristic time τ. In just a few multiples of the time τ, the final value is very nearly achieved, as the graph in [link](b) illustrates.

The characteristic time τ depends on only two factors, the inductance L and the resistance R. The greater the inductance L, the greater τ is, which makes sense since a large inductance is very effective in opposing change. The smaller the resistance R, the greater τ is. Again this makes sense, since a small resistance means a large final current and a greater change to get there. In both cases—large L and small R —more energy is stored in the inductor and more time is required to get it in and out.

When the switch in [link](a) is moved to position 2 and cuts the battery out of the circuit, the current drops because of energy dissipation by the resistor. But this is also not instantaneous, since the inductor opposes the decrease in current by inducing an emf in the same direction as the battery that drove

the current. Furthermore, there is a certain amount of energy, $(1/2)LI_0^2$, stored in the inductor, and it is dissipated at a finite rate. As the current approaches zero, the rate of decrease slows, since the energy dissipation rate is I^2R. Once again the behavior is exponential, and I is found to be

Equation:

$$I = I_0 e^{-t/\tau} \quad \text{(turning off)}.$$

(See [link](c).) In the first period of time $\tau = L/R$ after the switch is closed, the current falls to 0.368 of its initial value, since $I = I_0 e^{-1} = 0.368 I_0$. In each successive time τ, the current falls to 0.368 of the preceding value, and in a few multiples of τ, the current becomes very close to zero, as seen in the graph in [link](c).

Example:

Calculating Characteristic Time and Current in an *RL* Circuit

(a) What is the characteristic time constant for a 7.50 mH inductor in series with a $3.00\ \Omega$ resistor? (b) Find the current 5.00 ms after the switch is moved to position 2 to disconnect the battery, if it is initially 10.0 A.

Strategy for (a)

The time constant for an *RL* circuit is defined by $\tau = L/R$.

Solution for (a)

Entering known values into the expression for τ given in $\tau = L/R$ yields

Equation:

$$\tau = \frac{L}{R} = \frac{7.50\ \text{mH}}{3.00\ \Omega} = 2.50\ \text{ms}.$$

Discussion for (a)

This is a small but definitely finite time. The coil will be very close to its full current in about ten time constants, or about 25 ms.

Strategy for (b)

We can find the current by using $I = I_0 e^{-t/\tau}$, or by considering the decline in steps. Since the time is twice the characteristic time, we consider the process in steps.

Solution for (b)

In the first 2.50 ms, the current declines to 0.368 of its initial value, which is

Equation:

$$\begin{aligned} I &= 0.368 I_0 = (0.368)(10.0\text{ A}) \\ &= 3.68\text{ A at } t = 2.50\text{ ms.} \end{aligned}$$

After another 2.50 ms, or a total of 5.00 ms, the current declines to 0.368 of the value just found. That is,

Equation:

$$\begin{aligned} I\prime &= 0.368 I = (0.368)(3.68\text{ A}) \\ &= 1.35\text{ A at } t = 5.00\text{ ms.} \end{aligned}$$

Discussion for (b)

After another 5.00 ms has passed, the current will be 0.183 A (see [link]); so, although it does die out, the current certainly does not go to zero instantaneously.

In summary, when the voltage applied to an inductor is changed, the current also changes, *but the change in current lags the change in voltage in an RL circuit*. In Reactance, Inductive and Capacitive, we explore how an *RL* circuit behaves when a sinusoidal AC voltage is applied.

Section Summary

- When a series connection of a resistor and an inductor—an *RL* circuit —is connected to a voltage source, the time variation of the current is

  Equation:

  $$I = I_0(1 - e^{-t/\tau}) \quad \text{(turning on).}$$

  where $I_0 = V/R$ is the final current.

- The characteristic time constant τ is $\tau = \frac{L}{R}$, where L is the inductance and R is the resistance.
- In the first time constant τ, the current rises from zero to $0.632 I_0$, and 0.632 of the remainder in every subsequent time interval τ.
- When the inductor is shorted through a resistor, current decreases as **Equation:**

  $$I = I_0 e^{-t/\tau} \quad \text{(turning off)}.$$

  Here I_0 is the initial current.
- Current falls to $0.368 I_0$ in the first time interval τ, and 0.368 of the remainder toward zero in each subsequent time τ.

Problem Exercises

Exercise:

Problem:

If you want a characteristic *RL* time constant of 1.00 s, and you have a $500\ \Omega$ resistor, what value of self-inductance is needed?

Solution:

500 H

Exercise:

Problem:

Your *RL* circuit has a characteristic time constant of 20.0 ns, and a resistance of $5.00\ \text{M}\Omega$. (a) What is the inductance of the circuit? (b) What resistance would give you a 1.00 ns time constant, perhaps needed for quick response in an oscilloscope?

Exercise:

Problem:

A large superconducting magnet, used for magnetic resonance imaging, has a 50.0 H inductance. If you want current through it to be adjustable with a 1.00 s characteristic time constant, what is the minimum resistance of system?

Solution:

$50.0\ \Omega$

Exercise:

Problem:

Verify that after a time of 10.0 ms, the current for the situation considered in [link] will be 0.183 A as stated.

Exercise:

Problem:

Suppose you have a supply of inductors ranging from 1.00 nH to 10.0 H, and resistors ranging from $0.100\ \Omega$ to $1.00\ \text{M}\Omega$. What is the range of characteristic *RL* time constants you can produce by connecting a single resistor to a single inductor?

Solution:

1.00×10^{-18} s to 0.100 s

Exercise:

Problem:

(a) What is the characteristic time constant of a 25.0 mH inductor that has a resistance of $4.00\ \Omega$? (b) If it is connected to a 12.0 V battery, what is the current after 12.5 ms?

Exercise:

Problem:

What percentage of the final current I_0 flows through an inductor L in series with a resistor R, three time constants after the circuit is completed?

Solution:

95.0%

Exercise:

Problem:

The 5.00 A current through a 1.50 H inductor is dissipated by a $2.00\ \Omega$ resistor in a circuit like that in [link] with the switch in position 2. (a) What is the initial energy in the inductor? (b) How long will it take the current to decline to 5.00% of its initial value? (c) Calculate the average power dissipated, and compare it with the initial power dissipated by the resistor.

Exercise:

Problem:

(a) Use the exact exponential treatment to find how much time is required to bring the current through an 80.0 mH inductor in series with a $15.0\ \Omega$ resistor to 99.0% of its final value, starting from zero. (b) Compare your answer to the approximate treatment using integral numbers of τ. (c) Discuss how significant the difference is.

Solution:

(a) 24.6 ms

(b) 26.7 ms

(c) 9% difference, which is greater than the inherent uncertainty in the given parameters.

Exercise:

Problem:

(a) Using the exact exponential treatment, find the time required for the current through a 2.00 H inductor in series with a $0.500\ \Omega$ resistor to be reduced to 0.100% of its original value. (b) Compare your answer to the approximate treatment using integral numbers of τ. (c) Discuss how significant the difference is.

Glossary

characteristic time constant
: denoted by τ, of a particular series *RL* circuit is calculated by $\tau = \frac{L}{R}$, where L is the inductance and R is the resistance

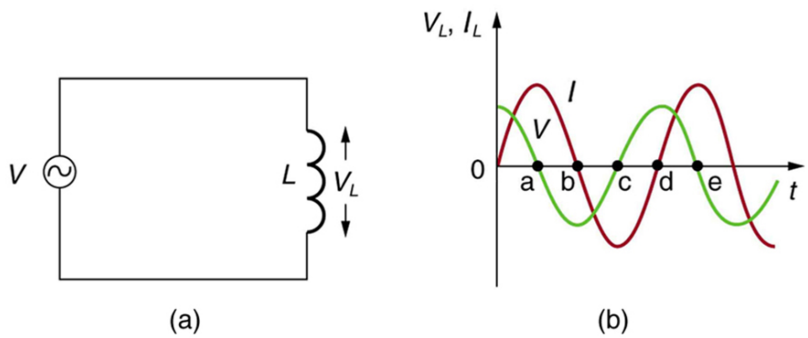
Reactance, Inductive and Capacitive

- Sketch voltage and current versus time in simple inductive, capacitive, and resistive circuits.
- Calculate inductive and capacitive reactance.
- Calculate current and/or voltage in simple inductive, capacitive, and resistive circuits.

Many circuits also contain capacitors and inductors, in addition to resistors and an AC voltage source. We have seen how capacitors and inductors respond to DC voltage when it is switched on and off. We will now explore how inductors and capacitors react to sinusoidal AC voltage.

Inductors and Inductive Reactance

Suppose an inductor is connected directly to an AC voltage source, as shown in [link]. It is reasonable to assume negligible resistance, since in practice we can make the resistance of an inductor so small that it has a negligible effect on the circuit. Also shown is a graph of voltage and current as functions of time.



(a) An AC voltage source in series with an inductor having negligible resistance. (b) Graph of current and voltage across the inductor as functions of time.

The graph in [link](b) starts with voltage at a maximum. Note that the current starts at zero and rises to its peak *after* the voltage that drives it, just as was the case when DC voltage was switched on in the preceding section. When the voltage becomes negative at point a, the current begins to decrease; it becomes zero at point b, where voltage is its most negative. The current then becomes negative, again following the voltage. The voltage becomes positive at point c and begins to make the current less negative. At point d, the current goes through zero just as the voltage reaches its positive peak to start another cycle. This behavior is summarized as follows:

Note:
AC Voltage in an Inductor
When a sinusoidal voltage is applied to an inductor, the voltage leads the current by one-fourth of a cycle, or by a 90º phase angle.

Current lags behind voltage, since inductors oppose change in current. Changing current induces a back emf $V = -L(\Delta I/\Delta t)$. This is considered to be an effective resistance of the inductor to AC. The rms current I through an inductor L is given by a version of Ohm's law:
Equation:

$$I = \frac{V}{X_L},$$

where V is the rms voltage across the inductor and X_L is defined to be
Equation:

$$X_L = 2\pi fL,$$

with f the frequency of the AC voltage source in hertz (An analysis of the circuit using Kirchhoff's loop rule and calculus actually produces this expression). X_L is called the **inductive reactance**, because the inductor

reacts to impede the current. X_L has units of ohms ($1\ \mathrm{H} = 1\ \Omega \cdot \mathrm{s}$, so that frequency times inductance has units of $(\mathrm{cycles/s})(\Omega \cdot \mathrm{s}) = \Omega$), consistent with its role as an effective resistance. It makes sense that X_L is proportional to L, since the greater the induction the greater its resistance to change. It is also reasonable that X_L is proportional to frequency f, since greater frequency means greater change in current. That is, $\Delta I/\Delta t$ is large for large frequencies (large f, small Δt). The greater the change, the greater the opposition of an inductor.

Example:

Calculating Inductive Reactance and then Current

(a) Calculate the inductive reactance of a 3.00 mH inductor when 60.0 Hz and 10.0 kHz AC voltages are applied. (b) What is the rms current at each frequency if the applied rms voltage is 120 V?

Strategy

The inductive reactance is found directly from the expression $X_L = 2\pi\mathrm{fL}$. Once X_L has been found at each frequency, Ohm's law as stated in the Equation $I = V/X_L$ can be used to find the current at each frequency.

Solution for (a)

Entering the frequency and inductance into Equation $X_L = 2\pi\mathrm{fL}$ gives

Equation:

$$X_L = 2\pi\mathrm{fL} = 6.28(60.0/\mathrm{s})(3.00\ \mathrm{mH}) = 1.13\ \Omega \text{ at } 60\ \mathrm{Hz}.$$

Similarly, at 10 kHz,

Equation:

$$X_L = 2\pi\mathrm{fL} = 6.28(1.00 \times 10^4/\mathrm{s})(3.00\ \mathrm{mH}) = 188\ \Omega \text{ at } 10\ \mathrm{kHz}.$$

Solution for (b)

The rms current is now found using the version of Ohm's law in Equation $I = V/X_L$, given the applied rms voltage is 120 V. For the first frequency, this yields

Equation:

$$I = \frac{V}{X_L} = \frac{120\ \text{V}}{1.13\ \Omega} = 106\ \text{A at 60 Hz}.$$

Similarly, at 10 kHz,

Equation:

$$I = \frac{V}{X_L} = \frac{120\ \text{V}}{188\ \Omega} = 0.637\ \text{A at 10 kHz}.$$

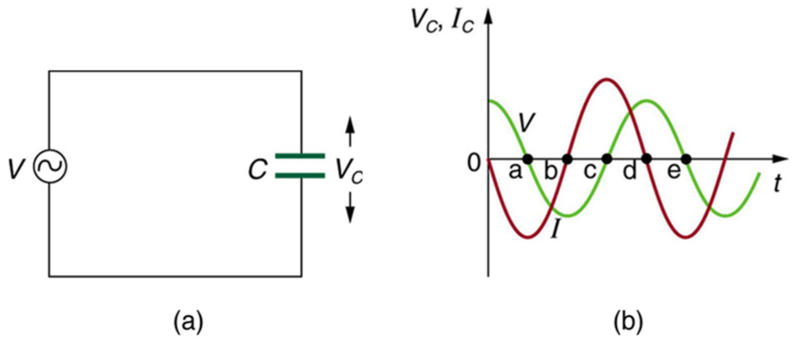
Discussion

The inductor reacts very differently at the two different frequencies. At the higher frequency, its reactance is large and the current is small, consistent with how an inductor impedes rapid change. Thus high frequencies are impeded the most. Inductors can be used to filter out high frequencies; for example, a large inductor can be put in series with a sound reproduction system or in series with your home computer to reduce high-frequency sound output from your speakers or high-frequency power spikes into your computer.

Note that although the resistance in the circuit considered is negligible, the AC current is not extremely large because inductive reactance impedes its flow. With AC, there is no time for the current to become extremely large.

Capacitors and Capacitive Reactance

Consider the capacitor connected directly to an AC voltage source as shown in [link]. The resistance of a circuit like this can be made so small that it has a negligible effect compared with the capacitor, and so we can assume negligible resistance. Voltage across the capacitor and current are graphed as functions of time in the figure.



(a) An AC voltage source in series with a capacitor *C* having negligible resistance. (b) Graph of current and voltage across the capacitor as functions of time.

The graph in [link] starts with voltage across the capacitor at a maximum. The current is zero at this point, because the capacitor is fully charged and halts the flow. Then voltage drops and the current becomes negative as the capacitor discharges. At point a, the capacitor has fully discharged ($Q=0$ on it) and the voltage across it is zero. The current remains negative between points a and b, causing the voltage on the capacitor to reverse. This is complete at point b, where the current is zero and the voltage has its most negative value. The current becomes positive after point b, neutralizing the charge on the capacitor and bringing the voltage to zero at point c, which allows the current to reach its maximum. Between points c and d, the current drops to zero as the voltage rises to its peak, and the process starts to repeat. Throughout the cycle, the voltage follows what the current is doing by one-fourth of a cycle:

Note:
AC Voltage in a Capacitor
When a sinusoidal voltage is applied to a capacitor, the voltage follows the current by one-fourth of a cycle, or by a 90° phase angle.

The capacitor is affecting the current, having the ability to stop it altogether when fully charged. Since an AC voltage is applied, there is an rms current, but it is limited by the capacitor. This is considered to be an effective resistance of the capacitor to AC, and so the rms current I in the circuit containing only a capacitor C is given by another version of Ohm's law to be

Equation:

$$I = \frac{V}{X_C},$$

where V is the rms voltage and X_C is defined (As with X_L, this expression for X_C results from an analysis of the circuit using Kirchhoff's rules and calculus) to be

Equation:

$$X_C = \frac{1}{2\pi f C},$$

where X_C is called the **capacitive reactance**, because the capacitor reacts to impede the current. X_C has units of ohms (verification left as an exercise for the reader). X_C is inversely proportional to the capacitance C; the larger the capacitor, the greater the charge it can store and the greater the current that can flow. It is also inversely proportional to the frequency f; the greater the frequency, the less time there is to fully charge the capacitor, and so it impedes current less.

Example:

Calculating Capacitive Reactance and then Current

(a) Calculate the capacitive reactance of a 5.00 mF capacitor when 60.0 Hz and 10.0 kHz AC voltages are applied. (b) What is the rms current if the applied rms voltage is 120 V?

Strategy

The capacitive reactance is found directly from the expression in $X_C = \frac{1}{2\pi fC}$. Once X_C has been found at each frequency, Ohm's law stated as $I = V/X_C$ can be used to find the current at each frequency.

Solution for (a)

Entering the frequency and capacitance into $X_C = \frac{1}{2\pi fC}$ gives

Equation:

$$\begin{aligned} X_C &= \frac{1}{2\pi fC} \\ &= \frac{1}{6.28(60.0/\text{s})(5.00\ \mu\text{F})} = 531\ \Omega \text{ at } 60\text{ Hz}. \end{aligned}$$

Similarly, at 10 kHz,

Equation:

$$\begin{aligned} X_C &= \frac{1}{2\pi fC} = \frac{1}{6.28(1.00\times10^4/\text{s})(5.00\ \mu\text{F})} \\ &= 3.18\ \Omega \text{ at } 10\text{ kHz} \end{aligned}.$$

Solution for (b)

The rms current is now found using the version of Ohm's law in $I = V/X_C$, given the applied rms voltage is 120 V. For the first frequency, this yields

Equation:

$$I = \frac{V}{X_C} = \frac{120\text{ V}}{531\ \Omega} = 0.226\text{ A at } 60\text{ Hz}.$$

Similarly, at 10 kHz,

Equation:

$$I = \frac{V}{X_C} = \frac{120\text{ V}}{3.18\ \Omega} = 37.7\text{ A at } 10\text{ kHz}.$$

Discussion

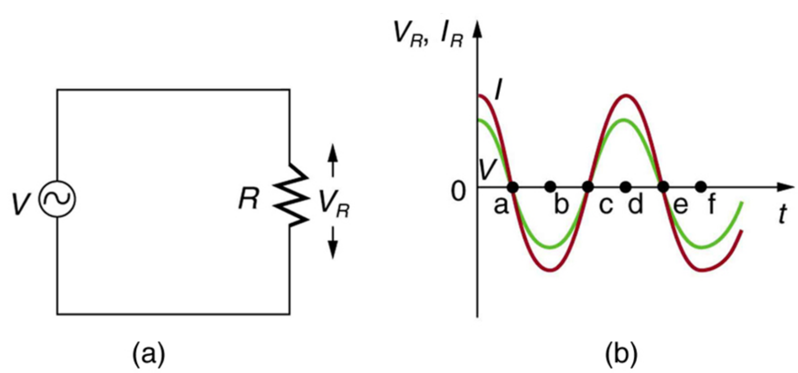
The capacitor reacts very differently at the two different frequencies, and in exactly the opposite way an inductor reacts. At the higher frequency, its reactance is small and the current is large. Capacitors favor change,

whereas inductors oppose change. Capacitors impede low frequencies the most, since low frequency allows them time to become charged and stop the current. Capacitors can be used to filter out low frequencies. For example, a capacitor in series with a sound reproduction system rids it of the 60 Hz hum.

Although a capacitor is basically an open circuit, there is an rms current in a circuit with an AC voltage applied to a capacitor. This is because the voltage is continually reversing, charging and discharging the capacitor. If the frequency goes to zero (DC), X_C tends to infinity, and the current is zero once the capacitor is charged. At very high frequencies, the capacitor's reactance tends to zero—it has a negligible reactance and does not impede the current (it acts like a simple wire). *Capacitors have the opposite effect on AC circuits that inductors have*.

Resistors in an AC Circuit

Just as a reminder, consider [link], which shows an AC voltage applied to a resistor and a graph of voltage and current versus time. The voltage and current are exactly *in phase* in a resistor. There is no frequency dependence to the behavior of plain resistance in a circuit:



(a) An AC voltage source in series with a resistor. (b) Graph of current and voltage

across the resistor as functions of time, showing them to be exactly in phase.

Note:
AC Voltage in a Resistor
When a sinusoidal voltage is applied to a resistor, the voltage is exactly in phase with the current—they have a 0º phase angle.

Section Summary

- For inductors in AC circuits, we find that when a sinusoidal voltage is applied to an inductor, the voltage leads the current by one-fourth of a cycle, or by a 90º phase angle.
- The opposition of an inductor to a change in current is expressed as a type of AC resistance.
- Ohm's law for an inductor is
  Equation:
  $$I=\frac{V}{X_L},$$
  where V is the rms voltage across the inductor.
- X_L is defined to be the inductive reactance, given by
  Equation:
  $$X_L=2\pi fL,$$
  with f the frequency of the AC voltage source in hertz.
- Inductive reactance X_L has units of ohms and is greatest at high frequencies.
- For capacitors, we find that when a sinusoidal voltage is applied to a capacitor, the voltage follows the current by one-fourth of a cycle, or

by a 90° phase angle.

- Since a capacitor can stop current when fully charged, it limits current and offers another form of AC resistance; Ohm's law for a capacitor is
  Equation:

$$I = \frac{V}{X_C},$$

  where V is the rms voltage across the capacitor.
- X_C is defined to be the capacitive reactance, given by
  Equation:

$$X_C = \frac{1}{2\pi f C}.$$

- X_C has units of ohms and is greatest at low frequencies.

Conceptual Questions

Exercise:

Problem:

Presbycusis is a hearing loss due to age that progressively affects higher frequencies. A hearing aid amplifier is designed to amplify all frequencies equally. To adjust its output for presbycusis, would you put a capacitor in series or parallel with the hearing aid's speaker? Explain.

Exercise:

Problem:

Would you use a large inductance or a large capacitance in series with a system to filter out low frequencies, such as the 100 Hz hum in a sound system? Explain.

Exercise:

Problem:

High-frequency noise in AC power can damage computers. Does the plug-in unit designed to prevent this damage use a large inductance or a large capacitance (in series with the computer) to filter out such high frequencies? Explain.

Exercise:

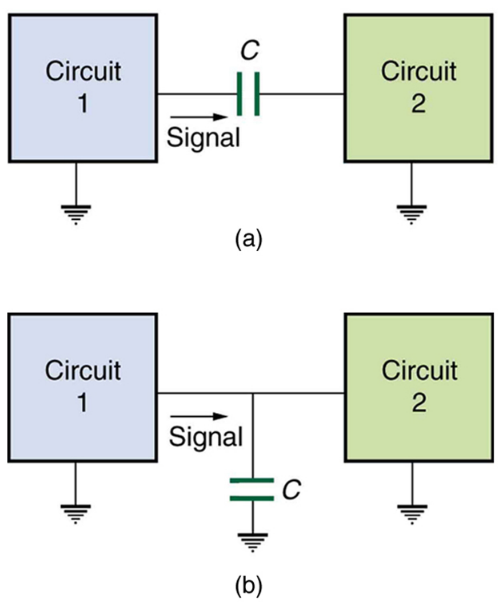
Problem:

Does inductance depend on current, frequency, or both? What about inductive reactance?

Exercise:

Problem:

Explain why the capacitor in [link](a) acts as a low-frequency filter between the two circuits, whereas that in [link](b) acts as a high-frequency filter.



Capacitors and inductors.

Capacitor with high frequency and low frequency.

Exercise:

Problem:

If the capacitors in [link] are replaced by inductors, which acts as a low-frequency filter and which as a high-frequency filter?

Problems & Exercises

Exercise:

Problem:

At what frequency will a 30.0 mH inductor have a reactance of $100\ \Omega$?

Solution:

531 Hz

Exercise:

Problem:

What value of inductance should be used if a $20.0\ \text{k}\Omega$ reactance is needed at a frequency of 500 Hz?

Exercise:

Problem:

What capacitance should be used to produce a $2.00\ \text{M}\Omega$ reactance at 60.0 Hz?

Solution:

1.33 nF

Exercise:

Problem:

At what frequency will an 80.0 mF capacitor have a reactance of $0.250\ \Omega$?

Exercise:

Problem:

(a) Find the current through a 0.500 H inductor connected to a 60.0 Hz, 480 V AC source. (b) What would the current be at 100 kHz?

Solution:

(a) 2.55 A

(b) 1.53 mA

Exercise:

Problem:

(a) What current flows when a 60.0 Hz, 480 V AC source is connected to a $0.250\ \mu\text{F}$ capacitor? (b) What would the current be at 25.0 kHz?

Exercise:

Problem:

A 20.0 kHz, 16.0 V source connected to an inductor produces a 2.00 A current. What is the inductance?

Solution:

63.7 µH

Exercise:

Problem:

A 20.0 Hz, 16.0 V source produces a 2.00 mA current when connected to a capacitor. What is the capacitance?

Exercise:

Problem:

(a) An inductor designed to filter high-frequency noise from power supplied to a personal computer is placed in series with the computer. What minimum inductance should it have to produce a $2.00\ \text{k}\Omega$ reactance for 15.0 kHz noise? (b) What is its reactance at 60.0 Hz?

Solution:

(a) 21.2 mH

(b) $8.00\ \Omega$

Exercise:

Problem:

The capacitor in [link](a) is designed to filter low-frequency signals, impeding their transmission between circuits. (a) What capacitance is needed to produce a $100\ \text{k}\Omega$ reactance at a frequency of 120 Hz? (b) What would its reactance be at 1.00 MHz? (c) Discuss the implications of your answers to (a) and (b).

Exercise:

Problem:

The capacitor in [link](b) will filter high-frequency signals by shorting them to earth/ground. (a) What capacitance is needed to produce a reactance of $10.0\ \text{m}\Omega$ for a 5.00 kHz signal? (b) What would its reactance be at 3.00 Hz? (c) Discuss the implications of your answers to (a) and (b).

Solution:

(a) 3.18 mF

(b) 16.7 Ω

Exercise:

Problem: Unreasonable Results

In a recording of voltages due to brain activity (an EEG), a 10.0 mV signal with a 0.500 Hz frequency is applied to a capacitor, producing a current of 100 mA. Resistance is negligible. (a) What is the capacitance? (b) What is unreasonable about this result? (c) Which assumption or premise is responsible?

Exercise:

Problem: Construct Your Own Problem

Consider the use of an inductor in series with a computer operating on 60 Hz electricity. Construct a problem in which you calculate the relative reduction in voltage of incoming high frequency noise compared to 60 Hz voltage. Among the things to consider are the acceptable series reactance of the inductor for 60 Hz power and the likely frequencies of noise coming through the power lines.

Glossary

inductive reactance
: the opposition of an inductor to a change in current; calculated by $X_L = 2\pi fL$

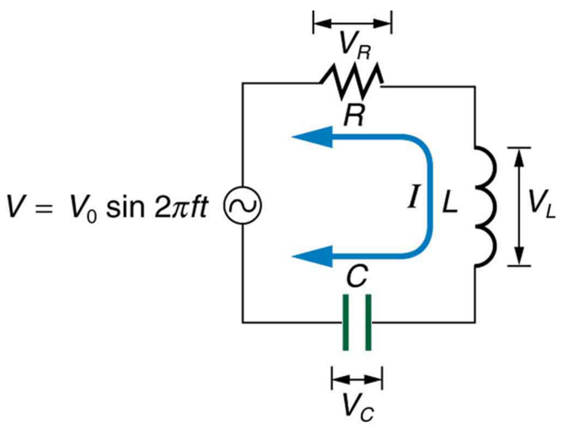
capacitive reactance
: the opposition of a capacitor to a change in current; calculated by $X_C = \frac{1}{2\pi fC}$

RLC Series AC Circuits

- Calculate the impedance, phase angle, resonant frequency, power, power factor, voltage, and/or current in a RLC series circuit.
- Draw the circuit diagram for an RLC series circuit.
- Explain the significance of the resonant frequency.

Impedance

When alone in an AC circuit, inductors, capacitors, and resistors all impede current. How do they behave when all three occur together? Interestingly, their individual resistances in ohms do not simply add. Because inductors and capacitors behave in opposite ways, they partially to totally cancel each other's effect. [link] shows an *RLC* series circuit with an AC voltage source, the behavior of which is the subject of this section. The crux of the analysis of an *RLC* circuit is the frequency dependence of X_L and X_C, and the effect they have on the phase of voltage versus current (established in the preceding section). These give rise to the frequency dependence of the circuit, with important "resonance" features that are the basis of many applications, such as radio tuners.



An *RLC* series circuit with an AC voltage source.

The combined effect of resistance R, inductive reactance X_L, and capacitive reactance X_C is defined to be **impedance**, an AC analogue to resistance in a DC circuit. Current, voltage, and impedance in an *RLC* circuit are related by an AC version of Ohm's law:

Equation:

$$I_0 = \frac{V_0}{Z} \text{ or } I_{\text{rms}} = \frac{V_{\text{rms}}}{Z}.$$

Here I_0 is the peak current, V_0 the peak source voltage, and Z is the impedance of the circuit. The units of impedance are ohms, and its effect on the circuit is as you might expect: the greater the impedance, the smaller the current. To get an expression for Z in terms of R, X_L, and X_C, we will now examine how the voltages across the various components are related to the source voltage. Those voltages are labeled V_R, V_L, and V_C in [link].

Conservation of charge requires current to be the same in each part of the circuit at all times, so that we can say the currents in R, L, and C are equal and in phase. But we know from the preceding section that the voltage across the inductor V_L leads the current by one-fourth of a cycle, the voltage across the capacitor V_C follows the current by one-fourth of a cycle, and the voltage across the resistor V_R is exactly in phase with the current. [link] shows these relationships in one graph, as well as showing the total voltage around the circuit $V = V_R + V_L + V_C$, where all four voltages are the instantaneous values. According to Kirchhoff's loop rule, the total voltage around the circuit V is also the voltage of the source.

You can see from [link] that while V_R is in phase with the current, V_L leads by 90º, and V_C follows by 90º. Thus V_L and V_C are 180º out of phase (crest to trough) and tend to cancel, although not completely unless they have the same magnitude. Since the peak voltages are not aligned (not in phase), the peak voltage V_0 of the source does *not* equal the sum of the peak voltages across R, L, and C. The actual relationship is

Equation:

$$V_0 = \sqrt{{V_{0R}}^2 + (V_{0L} - V_{0C})^2},$$

where V_{0R}, V_{0L}, and V_{0C} are the peak voltages across R, L, and C, respectively. Now, using Ohm's law and definitions from Reactance, Inductive and Capacitive, we substitute $V_0 = I_0 Z$ into the above, as well as $V_{0R} = I_0 R$, $V_{0L} = I_0 X_L$, and $V_{0C} = I_0 X_C$, yielding

Equation:

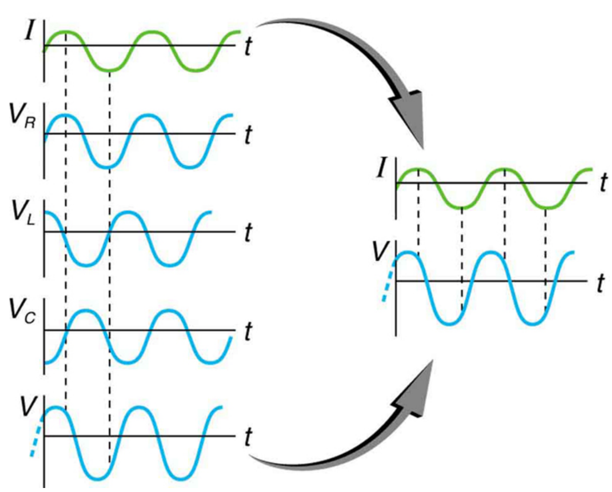
$$I_0 Z = \sqrt{{I_0}^2 R^2 + (I_0 X_L - I_0 X_C)^2} = I_0 \sqrt{R^2 + (X_L - X_C)^2}.$$

I_0 cancels to yield an expression for Z:

Equation:

$$Z = \sqrt{R^2 + (X_L - X_C)^2},$$

which is the impedance of an *RLC* series AC circuit. For circuits without a resistor, take $R = 0$; for those without an inductor, take $X_L = 0$; and for those without a capacitor, take $X_C = 0$.



This graph shows the relationships of the voltages in an *RLC* circuit to the current.

The voltages across the circuit elements add to equal the voltage of the source, which is seen to be out of phase with the current.

Example:

Calculating Impedance and Current

An *RLC* series circuit has a 40.0 Ω resistor, a 3.00 mH inductor, and a 5.00 μF capacitor. (a) Find the circuit's impedance at 60.0 Hz and 10.0 kHz, noting that these frequencies and the values for L and C are the same as in [link] and [link]. (b) If the voltage source has $V_{\text{rms}} = 120\text{ V}$, what is I_{rms} at each frequency?

Strategy

For each frequency, we use $Z = \sqrt{R^2 + (X_L - X_C)^2}$ to find the impedance and then Ohm's law to find current. We can take advantage of the results of the previous two examples rather than calculate the reactances again.

Solution for (a)

At 60.0 Hz, the values of the reactances were found in [link] to be $X_L = 1.13\ \Omega$ and in [link] to be $X_C = 531\ \Omega$. Entering these and the given 40.0 Ω for resistance into $Z = \sqrt{R^2 + (X_L - X_C)^2}$ yields

Equation:

$$\begin{aligned} Z &= \sqrt{R^2 + (X_L - X_C)^2} \\ &= \sqrt{(40.0\ \Omega)^2 + (1.13\ \Omega - 531\ \Omega)^2} \\ &= 531\ \Omega \text{ at } 60.0\text{ Hz}. \end{aligned}$$

Similarly, at 10.0 kHz, $X_L = 188\ \Omega$ and $X_C = 3.18\ \Omega$, so that

Equation:

$$\begin{aligned} Z &= \sqrt{(40.0\ \Omega)^2 + (188\ \Omega - 3.18\ \Omega)^2} \\ &= 190\ \Omega \text{ at } 10.0\ \text{kHz.} \end{aligned}$$

Discussion for (a)
In both cases, the result is nearly the same as the largest value, and the impedance is definitely not the sum of the individual values. It is clear that X_L dominates at high frequency and X_C dominates at low frequency.
Solution for (b)
The current I_{rms} can be found using the AC version of Ohm's law in Equation $I_{\text{rms}} = V_{\text{rms}}/Z$:
$I_{\text{rms}} = \frac{V_{\text{rms}}}{Z} = \frac{120\ \text{V}}{531\ \Omega} = 0.226\ \text{A at } 60.0\ \text{Hz}$
Finally, at 10.0 kHz, we find
$I_{\text{rms}} = \frac{V_{\text{rms}}}{Z} = \frac{120\ \text{V}}{190\ \Omega} = 0.633\ \text{A at } 10.0\ \text{kHz}$
Discussion for (a)
The current at 60.0 Hz is the same (to three digits) as found for the capacitor alone in [link]. The capacitor dominates at low frequency. The current at 10.0 kHz is only slightly different from that found for the inductor alone in [link]. The inductor dominates at high frequency.

Resonance in *RLC* Series AC Circuits

How does an *RLC* circuit behave as a function of the frequency of the driving voltage source? Combining Ohm's law, $I_{\text{rms}} = V_{\text{rms}}/Z$, and the expression for impedance Z from $Z = \sqrt{R^2 + (X_L - X_C)^2}$ gives
Equation:

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{\sqrt{R^2 + (X_L - X_C)^2}}.$$

The reactances vary with frequency, with X_L large at high frequencies and X_C large at low frequencies, as we have seen in three previous examples. At some intermediate frequency f_0, the reactances will be equal and cancel,

giving $Z = R$—this is a minimum value for impedance, and a maximum value for I_{rms} results. We can get an expression for f_0 by taking
Equation:

$$X_L = X_C.$$

Substituting the definitions of X_L and X_C,
Equation:

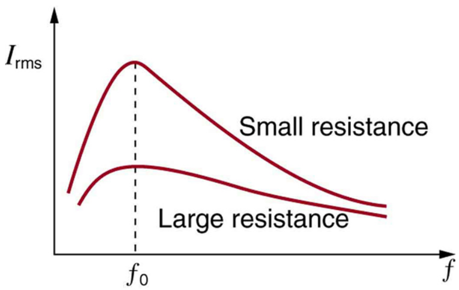
$$2\pi f_0 L = \frac{1}{2\pi f_0 C}.$$

Solving this expression for f_0 yields
Equation:

$$f_0 = \frac{1}{2\pi\sqrt{\mathrm{LC}}},$$

where f_0 is the **resonant frequency** of an *RLC* series circuit. This is also the *natural frequency* at which the circuit would oscillate if not driven by the voltage source. At f_0, the effects of the inductor and capacitor cancel, so that $Z = R$, and I_{rms} is a maximum.

Resonance in AC circuits is analogous to mechanical resonance, where resonance is defined to be a forced oscillation—in this case, forced by the voltage source—at the natural frequency of the system. The receiver in a radio is an *RLC* circuit that oscillates best at its f_0. A variable capacitor is often used to adjust f_0 to receive a desired frequency and to reject others. [link] is a graph of current as a function of frequency, illustrating a resonant peak in I_{rms} at f_0. The two curves are for two different circuits, which differ only in the amount of resistance in them. The peak is lower and broader for the higher-resistance circuit. Thus the higher-resistance circuit does not resonate as strongly and would not be as selective in a radio receiver, for example.



A graph of current versus frequency for two *RLC* series circuits differing only in the amount of resistance. Both have a resonance at f_0, but that for the higher resistance is lower and broader. The driving AC voltage source has a fixed amplitude V_0.

Example:
Calculating Resonant Frequency and Current
For the same *RLC* series circuit having a $40.0\ \Omega$ resistor, a 3.00 mH inductor, and a $5.00\ \mu\text{F}$ capacitor: (a) Find the resonant frequency. (b) Calculate I_{rms} at resonance if V_{rms} is 120 V.
Strategy
The resonant frequency is found by using the expression in $f_0 = \frac{1}{2\pi\sqrt{LC}}$.
The current at that frequency is the same as if the resistor alone were in the circuit.
Solution for (a)
Entering the given values for L and C into the expression given for f_0 in $f_0 = \frac{1}{2\pi\sqrt{LC}}$ yields

Equation:

$$\begin{aligned} f_0 &= \frac{1}{2\pi\sqrt{LC}} \\ &= \frac{1}{2\pi\sqrt{(3.00\times10^{-3}\text{ H})(5.00\times10^{-6}\text{ F})}} = 1.30\text{ kHz.} \end{aligned}$$

Discussion for (a)
We see that the resonant frequency is between 60.0 Hz and 10.0 kHz, the two frequencies chosen in earlier examples. This was to be expected, since the capacitor dominated at the low frequency and the inductor dominated at the high frequency. Their effects are the same at this intermediate frequency.
Solution for (b)
The current is given by Ohm's law. At resonance, the two reactances are equal and cancel, so that the impedance equals the resistance alone. Thus,
Equation:

$$I_{\text{rms}} = \frac{V_{\text{rms}}}{Z} = \frac{120\text{ V}}{40.0\ \Omega} = 3.00\text{ A.}$$

Discussion for (b)
At resonance, the current is greater than at the higher and lower frequencies considered for the same circuit in the preceding example.

Power in *RLC* Series AC Circuits

If current varies with frequency in an *RLC* circuit, then the power delivered to it also varies with frequency. But the average power is not simply current times voltage, as it is in purely resistive circuits. As was seen in [link], voltage and current are out of phase in an *RLC* circuit. There is a **phase angle** ϕ between the source voltage V and the current I, which can be found from
Equation:

$$\cos \phi = \frac{R}{Z}.$$

For example, at the resonant frequency or in a purely resistive circuit $Z = R$, so that $\cos \phi = 1$. This implies that $\phi = 0^{\circ}$ and that voltage and current are in phase, as expected for resistors. At other frequencies, average power is less than at resonance. This is both because voltage and current are out of phase and because I_{rms} is lower. The fact that source voltage and current are out of phase affects the power delivered to the circuit. It can be shown that the *average power* is

Equation:

$$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}} \cos \phi,$$

Thus $\cos \phi$ is called the **power factor**, which can range from 0 to 1. Power factors near 1 are desirable when designing an efficient motor, for example. At the resonant frequency, $\cos \phi = 1$.

Example:

Calculating the Power Factor and Power

For the same *RLC* series circuit having a 40.0 Ω resistor, a 3.00 mH inductor, a 5.00 μF capacitor, and a voltage source with a V_{rms} of 120 V: (a) Calculate the power factor and phase angle for $f = 60.0 \text{Hz}$. (b) What is the average power at 50.0 Hz? (c) Find the average power at the circuit's resonant frequency.

Strategy and Solution for (a)

The power factor at 60.0 Hz is found from

Equation:

$$\cos \phi = \frac{R}{Z}.$$

We know Z= 531 Ω from [link], so that

Equation:

$$\cos \phi = \frac{40.0\ \Omega}{531\ \Omega} = 0.0753 \text{ at } 60.0\text{ Hz}.$$

This small value indicates the voltage and current are significantly out of phase. In fact, the phase angle is

Equation:

$$\phi = \cos^{-1} 0.0753 = 85.7^{\text{o}} \text{ at } 60.0\text{ Hz}.$$

Discussion for (a)

The phase angle is close to 90º, consistent with the fact that the capacitor dominates the circuit at this low frequency (a pure *RC* circuit has its voltage and current 90º out of phase).

Strategy and Solution for (b)

The average power at 60.0 Hz is

Equation:

$$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}} \cos \phi.$$

I_{rms} was found to be 0.226 A in [link]. Entering the known values gives

Equation:

$$P_{\text{ave}} = (0.226\text{ A})(120\text{ V})(0.0753) = 2.04\text{ W at } 60.0\text{ Hz}.$$

Strategy and Solution for (c)

At the resonant frequency, we know $\cos \phi = 1$, and I_{rms} was found to be 6.00 A in [link]. Thus,

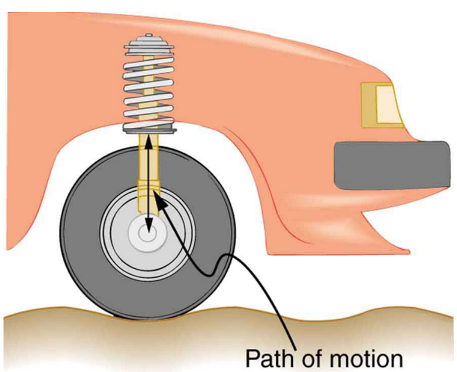
$P_{\text{ave}} = (3.00\text{ A})(120\text{ V})(1) = 360\text{ W}$ at resonance (1.30 kHz)

Discussion

Both the current and the power factor are greater at resonance, producing significantly greater power than at higher and lower frequencies.

Power delivered to an *RLC* series AC circuit is dissipated by the resistance alone. The inductor and capacitor have energy input and output but do not dissipate it out of the circuit. Rather they transfer energy back and forth to one another, with the resistor dissipating exactly what the voltage source

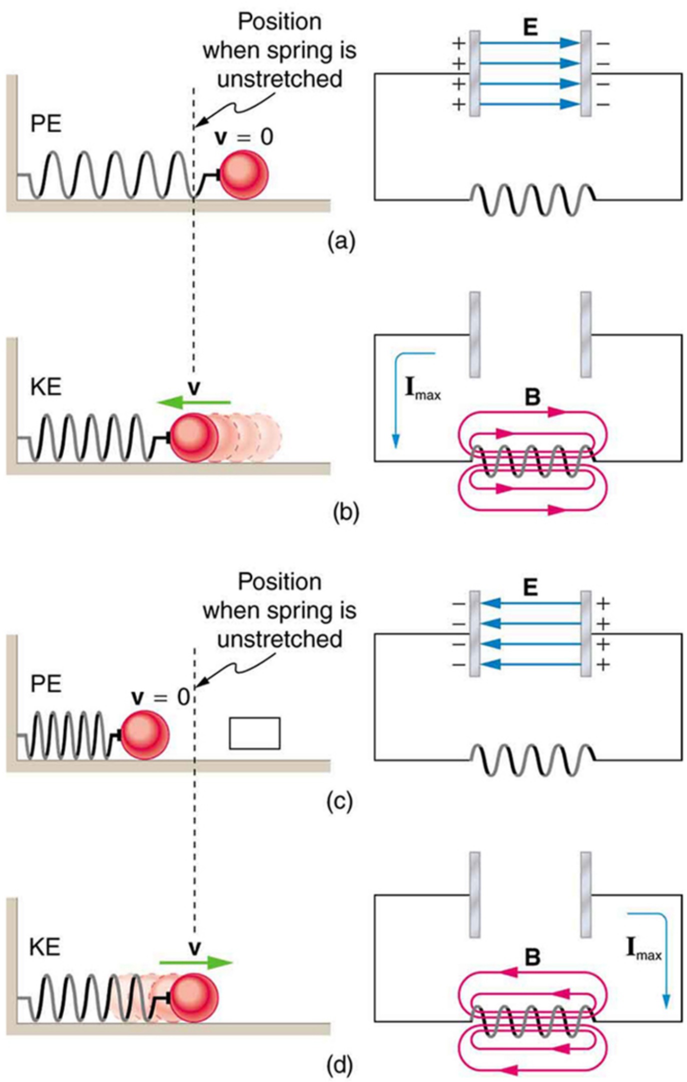
puts into the circuit. This assumes no significant electromagnetic radiation from the inductor and capacitor, such as radio waves. Such radiation can happen and may even be desired, as we will see in the next chapter on electromagnetic radiation, but it can also be suppressed as is the case in this chapter. The circuit is analogous to the wheel of a car driven over a corrugated road as shown in [link]. The regularly spaced bumps in the road are analogous to the voltage source, driving the wheel up and down. The shock absorber is analogous to the resistance damping and limiting the amplitude of the oscillation. Energy within the system goes back and forth between kinetic (analogous to maximum current, and energy stored in an inductor) and potential energy stored in the car spring (analogous to no current, and energy stored in the electric field of a capacitor). The amplitude of the wheels' motion is a maximum if the bumps in the road are hit at the resonant frequency.



The forced but damped motion of the wheel on the car spring is analogous to an *RLC* series AC circuit. The shock absorber damps the motion and dissipates energy, analogous to the resistance in an *RLC* circuit. The mass

and spring determine
the resonant
frequency.

A pure *LC* circuit with negligible resistance oscillates at f_0, the same resonant frequency as an *RLC* circuit. It can serve as a frequency standard or clock circuit—for example, in a digital wristwatch. With a very small resistance, only a very small energy input is necessary to maintain the oscillations. The circuit is analogous to a car with no shock absorbers. Once it starts oscillating, it continues at its natural frequency for some time. [link] shows the analogy between an *LC* circuit and a mass on a spring.



An *LC* circuit is analogous to a mass oscillating on a spring with no friction and no driving force. Energy moves back and forth between the inductor and capacitor, just as it moves from kinetic to potential in the mass-spring system.

Note:
PhET Explorations: Circuit Construction Kit (AC+DC), Virtual Lab
Build circuits with capacitors, inductors, resistors and AC or DC voltage sources, and inspect them using lab instruments such as voltmeters and ammeters.
https://phet.colorado.edu/sims/html/circuit-construction-kit-dc/latest/circuit-construction-kit-dc_en.html

Section Summary

- The AC analogy to resistance is impedance Z, the combined effect of resistors, inductors, and capacitors, defined by the AC version of Ohm's law:
  Equation:
  $$I_0 = \frac{V_0}{Z} \text{ or } I_{\text{rms}} = \frac{V_{\text{rms}}}{Z},$$
  where I_0 is the peak current and V_0 is the peak source voltage.
- Impedance has units of ohms and is given by $Z = \sqrt{R^2 + (X_L - X_C)^2}$.
- The resonant frequency f_0, at which $X_L = X_C$, is
  Equation:
  $$f_0 = \frac{1}{2\pi\sqrt{LC}}.$$
- In an AC circuit, there is a phase angle ϕ between source voltage V and the current I, which can be found from
  Equation:
  $$\cos\phi = \frac{R}{Z},$$
- $\phi = 0^\circ$ for a purely resistive circuit or an *RLC* circuit at resonance.

- The average power delivered to an *RLC* circuit is affected by the phase angle and is given by
  Equation:
  $$P_{\text{ave}} = I_{\text{rms}} V_{\text{rms}} \cos \phi,$$
  $\cos \phi$ is called the power factor, which ranges from 0 to 1.

Conceptual Questions

Exercise:

Problem:

Does the resonant frequency of an AC circuit depend on the peak voltage of the AC source? Explain why or why not.

Exercise:

Problem:

Suppose you have a motor with a power factor significantly less than 1. Explain why it would be better to improve the power factor as a method of improving the motor's output, rather than to increase the voltage input.

Problems & Exercises

Exercise:

Problem:

An *RL* circuit consists of a 40.0 Ω resistor and a 3.00 mH inductor. (a) Find its impedance Z at 60.0 Hz and 10.0 kHz. (b) Compare these values of Z with those found in [link] in which there was also a capacitor.

Solution:

(a) $40.02\ \Omega$ at 60.0 Hz, $193\ \Omega$ at 10.0 kHz

(b) At 60 Hz, with a capacitor, $\text{Z}=531\ \Omega$, over 13 times as high as without the capacitor. The capacitor makes a large difference at low frequencies. At 10 kHz, with a capacitor $\text{Z}=190\ \Omega$, about the same as without the capacitor. The capacitor has a smaller effect at high frequencies.

Exercise:

Problem:

An *RC* circuit consists of a $40.0\ \Omega$ resistor and a $5.00\ \mu\text{F}$ capacitor. (a) Find its impedance at 60.0 Hz and 10.0 kHz. (b) Compare these values of Z with those found in [link], in which there was also an inductor.

Exercise:

Problem:

An *LC* circuit consists of a $3.00\ \text{mH}$ inductor and a $5.00\ \mu F$ capacitor. (a) Find its impedance at 60.0 Hz and 10.0 kHz. (b) Compare these values of Z with those found in [link] in which there was also a resistor.

Solution:

(a) $529\ \Omega$ at 60.0 Hz, $185\ \Omega$ at 10.0 kHz

(b) These values are close to those obtained in [link] because at low frequency the capacitor dominates and at high frequency the inductor dominates. So in both cases the resistor makes little contribution to the total impedance.

Exercise:

Problem:

What is the resonant frequency of a 0.500 mH inductor connected to a 40.0 μF capacitor?

Exercise:

Problem:

To receive AM radio, you want an *RLC* circuit that can be made to resonate at any frequency between 500 and 1650 kHz. This is accomplished with a fixed 1.00 μH inductor connected to a variable capacitor. What range of capacitance is needed?

Solution:

9.30 nF to 101 nF

Exercise:

Problem:

Suppose you have a supply of inductors ranging from 1.00 nH to 10.0 H, and capacitors ranging from 1.00 pF to 0.100 F. What is the range of resonant frequencies that can be achieved from combinations of a single inductor and a single capacitor?

Exercise:

Problem:

What capacitance do you need to produce a resonant frequency of 1.00 GHz, when using an 8.00 nH inductor?

Solution:

3.17 pF

Exercise:

Problem:

What inductance do you need to produce a resonant frequency of 60.0 Hz, when using a 2.00 μF capacitor?

Exercise:

Problem:

The lowest frequency in the FM radio band is 88.0 MHz. (a) What inductance is needed to produce this resonant frequency if it is connected to a 2.50 pF capacitor? (b) The capacitor is variable, to allow the resonant frequency to be adjusted to as high as 108 MHz. What must the capacitance be at this frequency?

Solution:

(a) 1.31 μH

(b) 1.66 pF

Exercise:

Problem:

An *RLC* series circuit has a $2.50\ \Omega$ resistor, a $100\ \mu\text{H}$ inductor, and an $80.0\ \mu\text{F}$ capacitor.(a) Find the circuit's impedance at 120 Hz. (b) Find the circuit's impedance at 5.00 kHz. (c) If the voltage source has $V_{\text{rms}} = 5.60\ \text{V}$, what is I_{rms} at each frequency? (d) What is the resonant frequency of the circuit? (e) What is I_{rms} at resonance?

Exercise:

Problem:

An *RLC* series circuit has a $1.00\ \text{k}\Omega$ resistor, a $150\ \mu\text{H}$ inductor, and a 25.0 nF capacitor. (a) Find the circuit's impedance at 500 Hz. (b) Find the circuit's impedance at 7.50 kHz. (c) If the voltage source has $V_{\text{rms}} = 408\ \text{V}$, what is I_{rms} at each frequency? (d) What is the resonant frequency of the circuit? (e) What is I_{rms} at resonance?

Solution:

(a) $12.8\ \text{k}\Omega$

(b) $1.31\ \text{k}\Omega$

(c) 31.9 mA at 500 Hz, 312 mA at 7.50 kHz

(d) 82.2 kHz

(e) 0.408 A

Exercise:

Problem:

An *RLC* series circuit has a $2.50\ \Omega$ resistor, a $100\ \mu\text{H}$ inductor, and an $80.0\ \mu\text{F}$ capacitor. (a) Find the power factor at $f = 120\ \text{Hz}$. (b) What is the phase angle at 120 Hz? (c) What is the average power at 120 Hz? (d) Find the average power at the circuit's resonant frequency.

Exercise:

Problem:

An *RLC* series circuit has a $1.00\ \text{k}\Omega$ resistor, a $150\ \mu\text{H}$ inductor, and a 25.0 nF capacitor. (a) Find the power factor at $f = 7.50\ \text{Hz}$. (b) What is the phase angle at this frequency? (c) What is the average power at this frequency? (d) Find the average power at the circuit's resonant frequency.

Solution:

(a) 0.159

(b) 80.9º

(c) 26.4 W

(d) 166 W

Exercise:

Problem:

An *RLC* series circuit has a $200\ \Omega$ resistor and a 25.0 mH inductor. At 8000 Hz, the phase angle is 45.0^{o}. (a) What is the impedance? (b) Find the circuit's capacitance. (c) If $V_{\text{rms}} = 408\ \text{V}$ is applied, what is the average power supplied?

Exercise:

Problem: Referring to [link], find the average power at 10.0 kHz.

Solution:

16.0 W

Glossary

impedance
: the AC analogue to resistance in a DC circuit; it is the combined effect of resistance, inductive reactance, and capacitive reactance in the form $Z = \sqrt{R^2 + (X_L - X_C)^2}$

resonant frequency
: the frequency at which the impedance in a circuit is at a minimum, and also the frequency at which the circuit would oscillate if not driven by a voltage source; calculated by $f_0 = \frac{1}{2\pi\sqrt{LC}}$

phase angle
: denoted by ϕ, the amount by which the voltage and current are out of phase with each other in a circuit

power factor
: the amount by which the power delivered in the circuit is less than the theoretical maximum of the circuit due to voltage and current being out of phase; calculated by $\cos\phi$

Introduction to Wave Optics

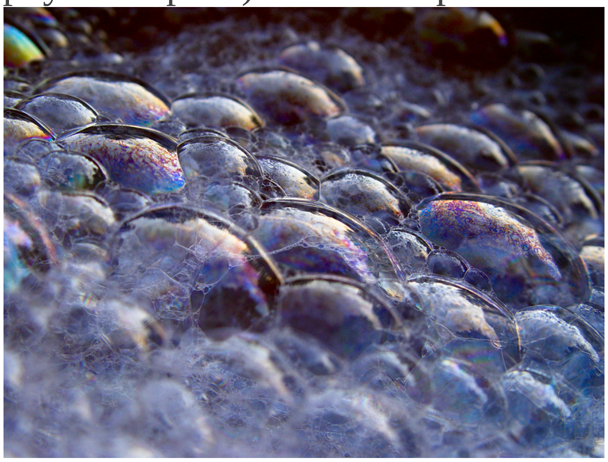
class="introduction"

The colors reflected by this compact disc vary with angle and are not caused by pigments. Colors such as these are direct evidence of the wave character of light. (credit: Infopro, Wikimedia Commons)

Examine a compact disc under white light, noting the colors observed and locations of the colors. Determine if the spectra are formed by diffraction from circular lines centered at the middle of the disc and, if so, what is their spacing. If not, determine the type of spacing. Also with the CD, explore the spectra of a few light sources, such as a candle flame, incandescent bulb, halogen light, and fluorescent light. Knowing the spacing of the rows of pits in the compact disc, estimate the maximum spacing that will allow the given number of megabytes of information to be stored.

If you have ever looked at the reds, blues, and greens in a sunlit soap bubble and wondered how straw-colored soapy water could produce them, you have hit upon one of the many phenomena that can only be explained by the wave character of light (see [link]). The same is true for the colors seen in an oil slick or in the light reflected from a compact disc. These and other interesting phenomena, such as the dispersion of white light into a rainbow of colors when passed through a narrow slit, cannot be explained fully by geometric optics. In these cases, light interacts with small objects and exhibits its wave characteristics. The branch of optics that considers the

behavior of light when it exhibits wave characteristics (particularly when it interacts with small objects) is called wave optics (sometimes called physical optics). It is the topic of this chapter.



These soap bubbles exhibit brilliant colors when exposed to sunlight. How are the colors produced if they are not pigments in the soap? (credit: Scott Robinson, Flickr)

The Wave Aspect of Light: Interference

- Discuss the wave character of light.
- Identify the changes when light enters a medium.

We know that visible light is the type of electromagnetic wave to which our eyes respond. Like all other electromagnetic waves, it obeys the equation

Equation:

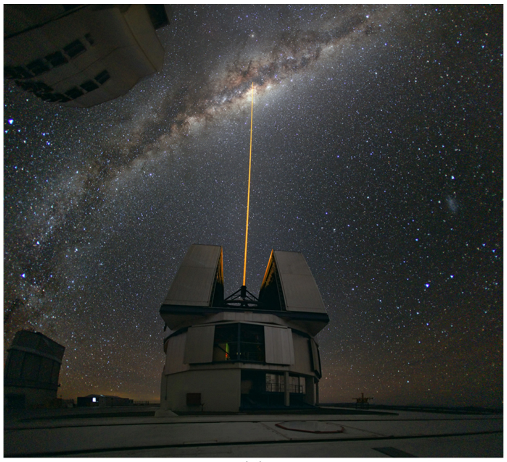
$$c = f\lambda,$$

where $c = 3 \times 10^8 \text{ m/s}$ is the speed of light in vacuum, f is the frequency of the electromagnetic waves, and λ is its wavelength. The range of visible wavelengths is approximately 380 to 760 nm. As is true for all waves, light travels in straight lines and acts like a ray when it interacts with objects several times as large as its wavelength. However, when it interacts with smaller objects, it displays its wave characteristics prominently. Interference is the hallmark of a wave, and in [link] both the ray and wave characteristics of light can be seen. The laser beam emitted by the observatory epitomizes a ray, traveling in a straight line. However, passing a pure-wavelength beam through vertical slits with a size close to the wavelength of the beam reveals the wave character of light, as the beam spreads out horizontally into a pattern of bright and dark regions caused by systematic constructive and destructive interference. Rather than spreading out, a ray would continue traveling straight ahead after passing through slits.

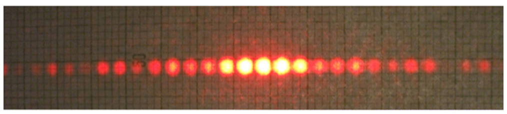
Note:

Making Connections: Waves

The most certain indication of a wave is interference. This wave characteristic is most prominent when the wave interacts with an object that is not large compared with the wavelength. Interference is observed for water waves, sound waves, light waves, and (as we will see in Special Relativity) for matter waves, such as electrons scattered from a crystal.



(a)



(b)

(a) The laser beam emitted by an observatory acts like a ray, traveling in a straight line. This laser beam is from the Paranal Observatory of the European Southern Observatory. (credit: Yuri Beletsky, European Southern Observatory) (b) A laser beam passing through a grid of vertical slits produces an interference pattern—characteristic of a wave. (credit: Shim'on and Slava Rybka, Wikimedia Commons)

Light has wave characteristics in various media as well as in a vacuum. When light goes from a vacuum to some medium, like water, its speed and

wavelength change, but its frequency f remains the same. (We can think of light as a forced oscillation that must have the frequency of the original source.) The speed of light in a medium is $v = c/n$, where n is its index of refraction. If we divide both sides of equation $c = f\lambda$ by n, we get $c/n = v = f\lambda/n$. This implies that $v = f\lambda_{\text{n}}$, where λ_{n} is the **wavelength in a medium** and that

Equation:

$$\lambda_{\text{n}} = \frac{\lambda}{n},$$

where λ is the wavelength in vacuum and n is the medium's index of refraction. Therefore, the wavelength of light is smaller in any medium than it is in vacuum. In water, for example, which has $n = 1.333$, the range of visible wavelengths is $(380\text{ nm})/1.333$ to $(760\text{ nm})/1.333$, or $\lambda_{\text{n}} = 285\text{ to }570\text{ nm}$. Although wavelengths change while traveling from one medium to another, colors do not, since colors are associated with frequency.

Section Summary

- Wave optics is the branch of optics that must be used when light interacts with small objects or whenever the wave characteristics of light are considered.
- Wave characteristics are those associated with interference and diffraction.
- Visible light is the type of electromagnetic wave to which our eyes respond and has a wavelength in the range of 380 to 760 nm.
- Like all EM waves, the following relationship is valid in vacuum: $c = f\lambda$, where $c = 3 \times 10^8\text{ m/s}$ is the speed of light, f is the frequency of the electromagnetic wave, and λ is its wavelength in vacuum.
- The wavelength λ_{n} of light in a medium with index of refraction n is $\lambda_{\text{n}} = \lambda/n$. Its frequency is the same as in vacuum.

Conceptual Questions

Exercise:

Problem:

What type of experimental evidence indicates that light is a wave?

Exercise:

Problem:

Give an example of a wave characteristic of light that is easily observed outside the laboratory.

Problems & Exercises

Exercise:

Problem:

Show that when light passes from air to water, its wavelength decreases to 0.750 times its original value.

Solution:

$1/1.333 = 0.750$

Exercise:

Problem:

Find the range of visible wavelengths of light in crown glass.

Exercise:

Problem:

What is the index of refraction of a material for which the wavelength of light is 0.671 times its value in a vacuum? Identify the likely substance.

Solution:

1.49, Polystyrene

Exercise:

Problem:

Analysis of an interference effect in a clear solid shows that the wavelength of light in the solid is 329 nm. Knowing this light comes from a He-Ne laser and has a wavelength of 633 nm in air, is the substance zircon or diamond?

Exercise:

Problem:

What is the ratio of thicknesses of crown glass and water that would contain the same number of wavelengths of light?

Solution:

0.877 glass to water

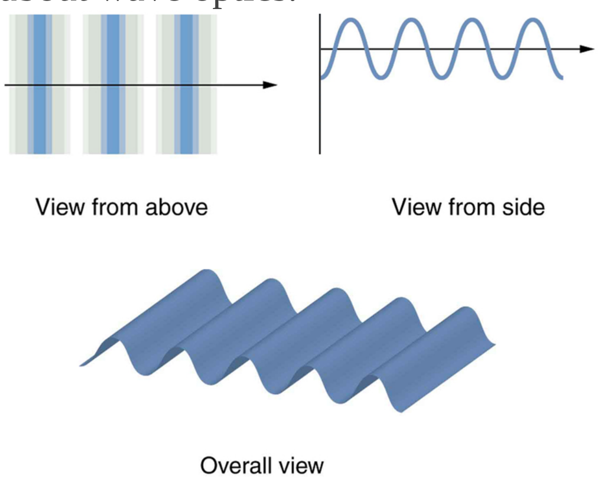
Glossary

wavelength in a medium
: $\lambda_{\mathrm{n}} = \lambda / n$, where λ is the wavelength in vacuum, and n is the index of refraction of the medium

Huygens's Principle: Diffraction

- Discuss the propagation of transverse waves.
- Discuss Huygens’s principle.
- Explain the bending of light.

[link] shows how a transverse wave looks as viewed from above and from the side. A light wave can be imagined to propagate like this, although we do not actually see it wiggling through space. From above, we view the wavefronts (or wave crests) as we would by looking down on the ocean waves. The side view would be a graph of the electric or magnetic field. The view from above is perhaps the most useful in developing concepts about wave optics.



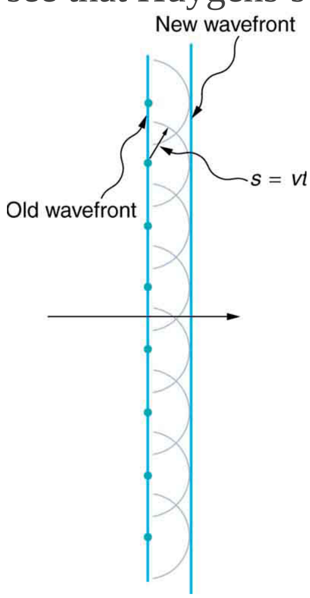
A transverse wave, such as an electromagnetic wave like light, as viewed from above and from the side. The direction of propagation is perpendicular to the wavefronts (or wave crests) and is represented by an arrow like a ray.

The Dutch scientist Christiaan Huygens (1629–1695) developed a useful technique for determining in detail how and where waves propagate.

Starting from some known position, **Huygens's principle** states that:

Every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets.

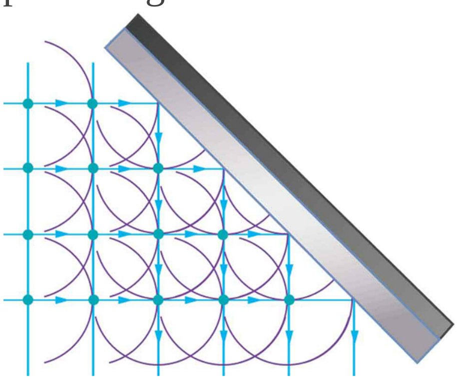
[link] shows how Huygens's principle is applied. A wavefront is the long edge that moves, for example, the crest or the trough. Each point on the wavefront emits a semicircular wave that moves at the propagation speed v. These are drawn at a time t later, so that they have moved a distance $s = \text{vt}$. The new wavefront is a line tangent to the wavelets and is where we would expect the wave to be a time t later. Huygens's principle works for all types of waves, including water waves, sound waves, and light waves. We will find it useful not only in describing how light waves propagate, but also in explaining the laws of reflection and refraction. In addition, we will see that Huygens's principle tells us how and where light rays interfere.



Huygens's principle applied to a straight wavefront. Each point on the wavefront

emits a semicircular wavelet that moves a distance $s = vt$. The new wavefront is a line tangent to the wavelets.

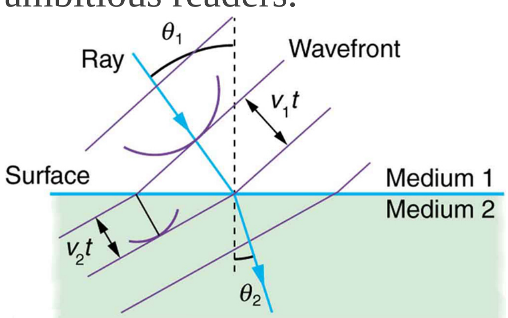
[link] shows how a mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection. As the wavefront strikes the mirror, wavelets are first emitted from the left part of the mirror and then the right. The wavelets closer to the left have had time to travel farther, producing a wavefront traveling in the direction shown.



Huygens’s principle applied to a straight wavefront striking a mirror. The wavelets shown were emitted as each point on the wavefront struck the mirror. The tangent to these wavelets shows

that the new wavefront has been reflected at an angle equal to the incident angle. The direction of propagation is perpendicular to the wavefront, as shown by the downward-pointing arrows.

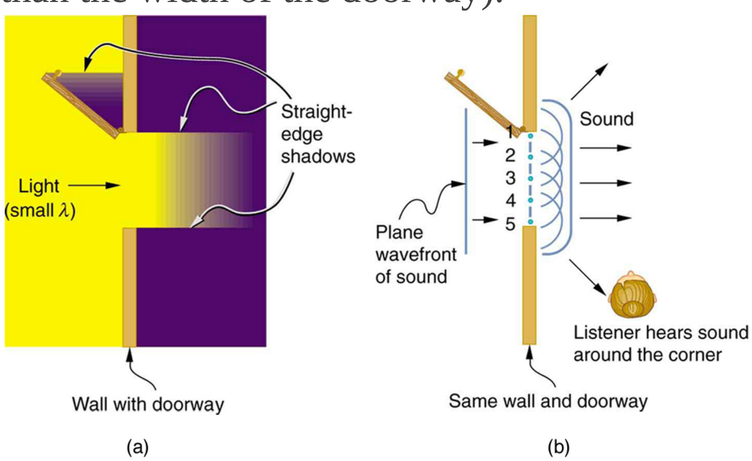
The law of refraction can be explained by applying Huygens's principle to a wavefront passing from one medium to another (see [link]). Each wavelet in the figure was emitted when the wavefront crossed the interface between the media. Since the speed of light is smaller in the second medium, the waves do not travel as far in a given time, and the new wavefront changes direction as shown. This explains why a ray changes direction to become closer to the perpendicular when light slows down. Snell's law can be derived from the geometry in [link], but this is left as an exercise for ambitious readers.



Huygens's principle applied to a straight wavefront traveling from one medium to another where its speed is less. The ray bends toward the perpendicular, since the

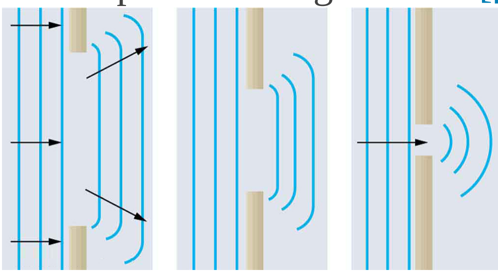
wavelets have a lower speed in the second medium.

What happens when a wave passes through an opening, such as light shining through an open door into a dark room? For light, we expect to see a sharp shadow of the doorway on the floor of the room, and we expect no light to bend around corners into other parts of the room. When sound passes through a door, we expect to hear it everywhere in the room and, thus, expect that sound spreads out when passing through such an opening (see [link]). What is the difference between the behavior of sound waves and light waves in this case? The answer is that light has very short wavelengths and acts like a ray. Sound has wavelengths on the order of the size of the door and bends around corners (for frequency of 1000 Hz, $\lambda = c/f = (330\text{ m/s})/(1000\text{ s}^{-1}) = 0.33\text{ m}$, about three times smaller than the width of the doorway).



(a) Light passing through a doorway makes a sharp outline on the floor. Since light’s wavelength is very small compared with the size of the door, it acts like a ray. (b) Sound waves bend into all parts of the room, a wave effect, because their wavelength is similar to the size of the door.

If we pass light through smaller openings, often called slits, we can use Huygens’s principle to see that light bends as sound does (see [link]). The bending of a wave around the edges of an opening or an obstacle is called **diffraction**. Diffraction is a wave characteristic and occurs for all types of waves. If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Thus the horizontal diffraction of the laser beam after it passes through slits in [link] is evidence that light is a wave.



Huygens’s principle applied to a straight wavefront striking an opening. The edges of the wavefront bend after passing through the opening, a process called diffraction. The amount of bending is more extreme for a small opening, consistent with the fact that wave characteristics are most noticeable for interactions with objects about the same size as the wavelength.

Section Summary

- An accurate technique for determining how and where waves propagate is given by Huygens's principle: Every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets.
- Diffraction is the bending of a wave around the edges of an opening or other obstacle.

Conceptual Questions

Exercise:

Problem:

How do wave effects depend on the size of the object with which the wave interacts? For example, why does sound bend around the corner of a building while light does not?

Exercise:

Problem:

Under what conditions can light be modeled like a ray? Like a wave?

Exercise:

Problem:

Go outside in the sunlight and observe your shadow. It has fuzzy edges even if you do not. Is this a diffraction effect? Explain.

Exercise:

Problem:

Why does the wavelength of light decrease when it passes from vacuum into a medium? State which attributes change and which stay the same and, thus, require the wavelength to decrease.

Exercise:

Problem: Does Huygens's principle apply to all types of waves?

Glossary

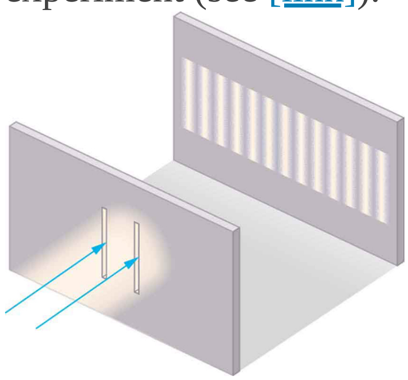
diffraction
: the bending of a wave around the edges of an opening or an obstacle

Huygens's principle
: every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets

Young’s Double Slit Experiment

- Explain the phenomena of interference.
- Define constructive interference for a double slit and destructive interference for a double slit.

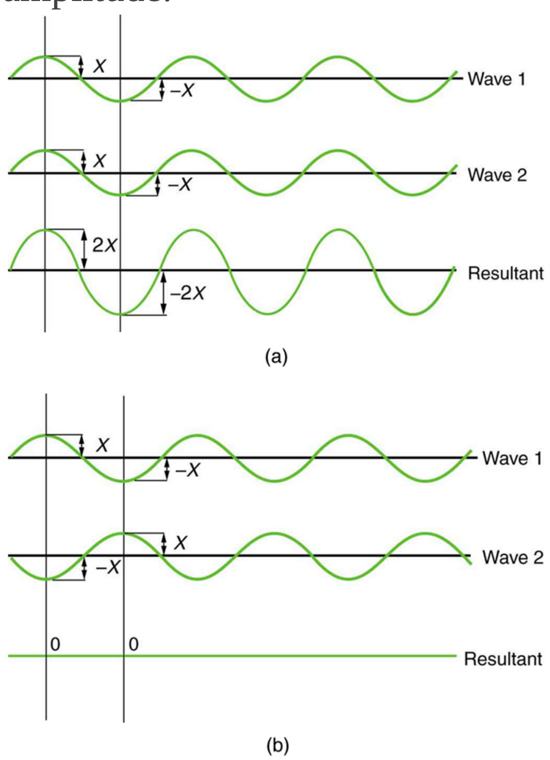
Although Christiaan Huygens thought that light was a wave, Isaac Newton did not. Newton felt that there were other explanations for color, and for the interference and diffraction effects that were observable at the time. Owing to Newton’s tremendous stature, his view generally prevailed. The fact that Huygens’s principle worked was not considered evidence that was direct enough to prove that light is a wave. The acceptance of the wave character of light came many years later when, in 1801, the English physicist and physician Thomas Young (1773–1829) did his now-classic double slit experiment (see [link]).



Young’s double slit experiment. Here pure-wavelength light sent through a pair of vertical slits is diffracted into a pattern on the screen of numerous vertical lines spread out horizontally. Without diffraction and interference, the light would

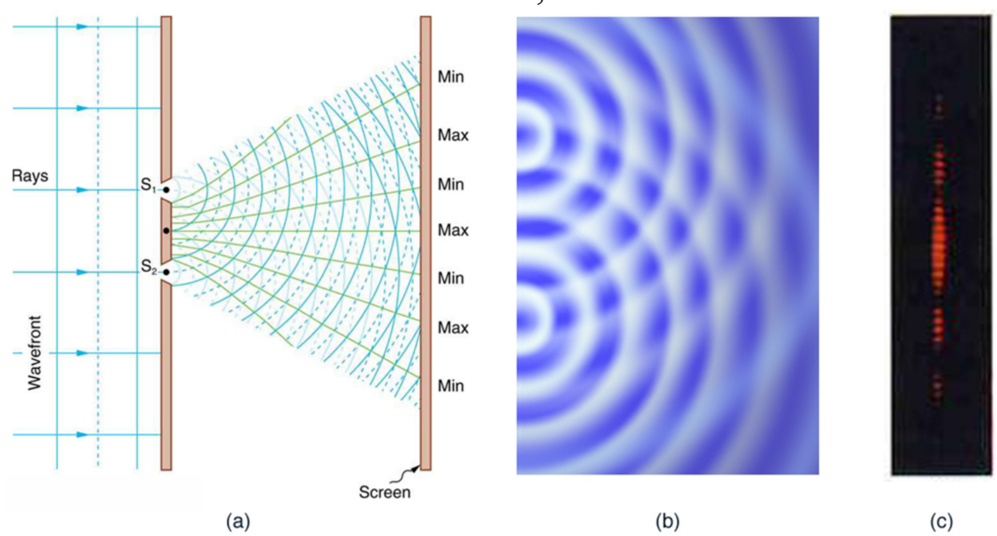
simply make two
lines on the screen.

Why do we not ordinarily observe wave behavior for light, such as observed in Young's double slit experiment? First, light must interact with something small, such as the closely spaced slits used by Young, to show pronounced wave effects. Furthermore, Young first passed light from a single source (the Sun) through a single slit to make the light somewhat coherent. By **coherent**, we mean waves are in phase or have a definite phase relationship. **Incoherent** means the waves have random phase relationships. Why did Young then pass the light through a double slit? The answer to this question is that two slits provide two coherent light sources that then interfere constructively or destructively. Young used sunlight, where each wavelength forms its own pattern, making the effect more difficult to see. We illustrate the double slit experiment with monochromatic (single λ) light to clarify the effect. [link] shows the pure constructive and destructive interference of two waves having the same wavelength and amplitude.



The amplitudes of waves add. (a) Pure constructive interference is obtained when identical waves are in phase. (b) Pure destructive interference occurs when identical waves are exactly out of phase, or shifted by half a wavelength.

When light passes through narrow slits, it is diffracted into semicircular waves, as shown in [link](a). Pure constructive interference occurs where the waves are crest to crest or trough to trough. Pure destructive interference occurs where they are crest to trough. The light must fall on a screen and be scattered into our eyes for us to see the pattern. An analogous pattern for water waves is shown in [link](b). Note that regions of constructive and destructive interference move out from the slits at well-defined angles to the original beam. These angles depend on wavelength and the distance between the slits, as we shall see below.

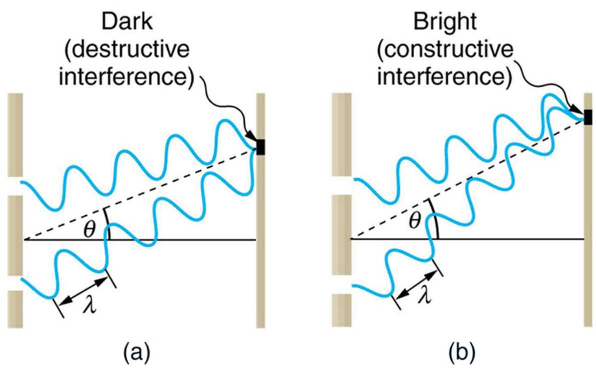


Double slits produce two coherent sources of waves that interfere. (a) Light spreads out (diffracts) from each slit, because the slits are narrow. These waves overlap and

interfere constructively (bright lines) and destructively (dark regions). We can only see this if the light falls onto a screen and is scattered into our eyes. (b) Double slit interference pattern for water waves are nearly identical to that for light. Wave action is greatest in regions of constructive interference and least in regions of destructive interference. (c) When light that has passed through double slits falls on a screen, we see a pattern such as this. (credit: PASCO)

To understand the double slit interference pattern, we consider how two waves travel from the slits to the screen, as illustrated in [link]. Each slit is a different distance from a given point on the screen. Thus different numbers of wavelengths fit into each path. Waves start out from the slits in phase (crest to crest), but they may end up out of phase (crest to trough) at the screen if the paths differ in length by half a wavelength, interfering destructively as shown in [link](a). If the paths differ by a whole wavelength, then the waves arrive in phase (crest to crest) at the screen, interfering constructively as shown in [link](b). More generally, if the paths taken by the two waves differ by any half-integral number of wavelengths [$(1/2)\lambda$, $(3/2)\lambda$, $(5/2)\lambda$, etc.], then destructive interference occurs. Similarly, if the paths taken by the two waves differ by any integral number of wavelengths (λ, 2λ, 3λ, etc.), then constructive interference occurs.

Note:
Take-Home Experiment: Using Fingers as Slits
Look at a light, such as a street lamp or incandescent bulb, through the narrow gap between two fingers held close together. What type of pattern do you see? How does it change when you allow the fingers to move a little farther apart? Is it more distinct for a monochromatic source, such as the yellow light from a sodium vapor lamp, than for an incandescent bulb?



Waves follow different paths from the slits to a common point on a screen. (a) Destructive interference occurs here, because one path is a half wavelength longer than the other. The waves start in phase but arrive out of phase. (b) Constructive interference occurs here because one path is a whole wavelength longer than the other. The waves start out and arrive in phase.

[link] shows how to determine the path length difference for waves traveling from two slits to a common point on a screen. If the screen is a large distance away compared with the distance between the slits, then the angle θ between the path and a line from the slits to the screen (see the figure) is nearly the same for each path. The difference between the paths is shown in the figure; simple trigonometry shows it to be $d\sin\theta$, where d is the distance between the slits. To obtain **constructive interference for a double slit**, the path length difference must be an integral multiple of the wavelength, or

Equation:

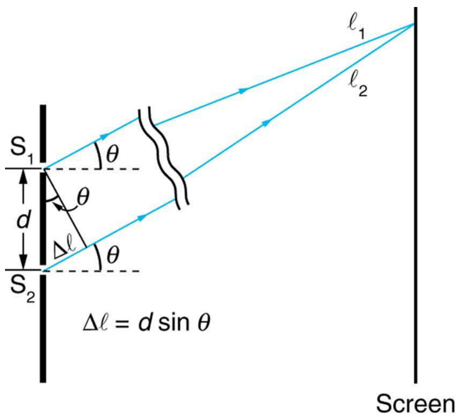
$$d\ \sin\theta = m\lambda,\ \text{for}\ m = 0, 1, -1, 2, -2, \ldots\ \text{(constructive)}.$$

Similarly, to obtain **destructive interference for a double slit**, the path length difference must be a half-integral multiple of the wavelength, or

Equation:

$$d\ \sin\theta = \left(m + \frac{1}{2}\right)\lambda,\ \text{for}\ m = 0, 1,\ -1, 2,\ -2,\ \ldots\ \ (\text{destructive}),$$

where λ is the wavelength of the light, d is the distance between slits, and θ is the angle from the original direction of the beam as discussed above. We call m the **order** of the interference. For example, $m = 4$ is fourth-order interference.



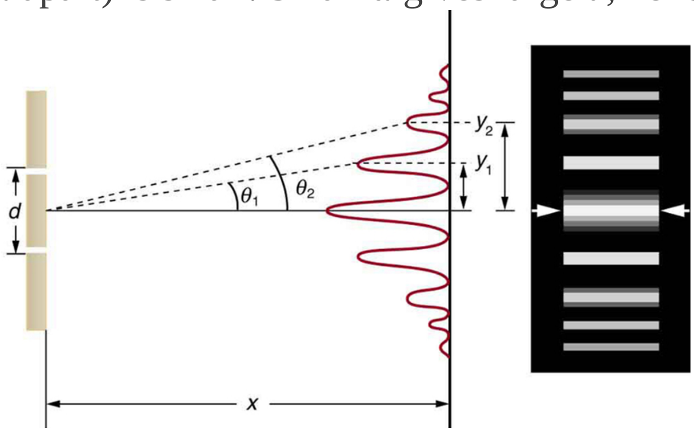
The paths from each slit to a common point on the screen differ by an amount $d\ \sin\theta$, assuming the distance to the screen is much greater than the distance between slits (not to scale here).

The equations for double slit interference imply that a series of bright and dark lines are formed. For vertical slits, the light spreads out horizontally on

either side of the incident beam into a pattern called interference fringes, illustrated in [link]. The intensity of the bright fringes falls off on either side, being brightest at the center. The closer the slits are, the more is the spreading of the bright fringes. We can see this by examining the equation
Equation:

$$d\ \sin\theta = m\lambda\text{, for } m = 0, 1, \ -1, 2, \ -2, \ \ldots.$$

For fixed λ and m, the smaller d is, the larger θ must be, since $\sin\theta = m\lambda/\ d$. This is consistent with our contention that wave effects are most noticeable when the object the wave encounters (here, slits a distance d apart) is small. Small d gives large θ, hence a large effect.



The interference pattern for a double slit has an intensity that falls off with angle. The photograph shows multiple bright and dark lines, or fringes, formed by light passing through a double slit.

Example:
Finding a Wavelength from an Interference Pattern

Suppose you pass light from a He-Ne laser through two slits separated by 0.0100 mm and find that the third bright line on a screen is formed at an angle of 10.95° relative to the incident beam. What is the wavelength of the light?

Strategy

The third bright line is due to third-order constructive interference, which means that $m = 3$. We are given $d = 0.0100$ mm and $\theta = 10.95^{\circ}$. The wavelength can thus be found using the equation $d \ \sin \theta = m\lambda$ for constructive interference.

Solution

The equation is $d \ \sin \theta = m\lambda$. Solving for the wavelength λ gives

Equation:

$$\lambda = \frac{d \ \sin \theta}{m}.$$

Substituting known values yields

Equation:

$$\begin{aligned} \lambda &= \frac{(0.0100 \text{ mm})(\sin 10.95^{\circ})}{3} \\ &= 6.33 \times 10^{-4} \text{ mm} = 633 \text{ nm.} \end{aligned}$$

Discussion

To three digits, this is the wavelength of light emitted by the common He-Ne laser. Not by coincidence, this red color is similar to that emitted by neon lights. More important, however, is the fact that interference patterns can be used to measure wavelength. Young did this for visible wavelengths. This analytical technique is still widely used to measure electromagnetic spectra. For a given order, the angle for constructive interference increases with λ, so that spectra (measurements of intensity versus wavelength) can be obtained.

Example:

Calculating Highest Order Possible

Interference patterns do not have an infinite number of lines, since there is a limit to how big m can be. What is the highest-order constructive interference possible with the system described in the preceding example?

Strategy and Concept

The equation $d\ \sin\theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$) describes constructive interference. For fixed values of d and λ, the larger m is, the larger $\sin\theta$ is. However, the maximum value that $\sin\theta$ can have is 1, for an angle of 90º. (Larger angles imply that light goes backward and does not reach the screen at all.) Let us find which m corresponds to this maximum diffraction angle.

Solution

Solving the equation $d \sin\theta = m\lambda$ for m gives

Equation:

$$m = \frac{d \sin\theta}{\lambda}.$$

Taking $\sin\theta = 1$ and substituting the values of d and λ from the preceding example gives

Equation:

$$m = \frac{(0.0100\text{ mm})(1)}{633\text{ nm}} \approx 15.8.$$

Therefore, the largest integer m can be is 15, or

Equation:

$$m = 15.$$

Discussion

The number of fringes depends on the wavelength and slit separation. The number of fringes will be very large for large slit separations. However, if the slit separation becomes much greater than the wavelength, the intensity of the interference pattern changes so that the screen has two bright lines cast by the slits, as expected when light behaves like a ray. We also note that the fringes get fainter further away from the center. Consequently, not all 15 fringes may be observable.

Section Summary

- Young’s double slit experiment gave definitive proof of the wave character of light.
- An interference pattern is obtained by the superposition of light from two slits.
- There is constructive interference when $d\sin\theta = m\lambda \;(\text{for } m = 0, 1, -1, 2, -2, \ldots)$, where d is the distance between the slits, θ is the angle relative to the incident direction, and m is the order of the interference.
- There is destructive interference when $d\ \sin\theta = \left(m + \frac{1}{2}\right)\lambda \;(\text{for } m = 0, 1, -1, 2, -2, \ldots)$.

Conceptual Questions

Exercise:

Problem:

Young’s double slit experiment breaks a single light beam into two sources. Would the same pattern be obtained for two independent sources of light, such as the headlights of a distant car? Explain.

Exercise:

Problem:

Suppose you use the same double slit to perform Young’s double slit experiment in air and then repeat the experiment in water. Do the angles to the same parts of the interference pattern get larger or smaller? Does the color of the light change? Explain.

Exercise:

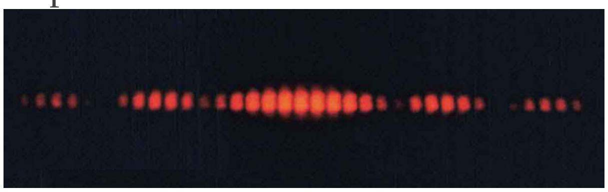
Problem:

Is it possible to create a situation in which there is only destructive interference? Explain.

Exercise:

Problem:

[link] shows the central part of the interference pattern for a pure wavelength of red light projected onto a double slit. The pattern is actually a combination of single slit and double slit interference. Note that the bright spots are evenly spaced. Is this a double slit or single slit characteristic? Note that some of the bright spots are dim on either side of the center. Is this a single slit or double slit characteristic? Which is smaller, the slit width or the separation between slits? Explain your responses.



This double slit interference pattern also shows signs of single slit interference. (credit: PASCO)

Problems & Exercises

Exercise:

Problem:

At what angle is the first-order maximum for 450-nm wavelength blue light falling on double slits separated by 0.0500 mm?

Solution:

0.516°

Exercise:

Problem:

Calculate the angle for the third-order maximum of 580-nm wavelength yellow light falling on double slits separated by 0.100 mm.

Exercise:

Problem:

What is the separation between two slits for which 610-nm orange light has its first maximum at an angle of 30.0^{o}?

Solution:

$1.22 \times 10^{-6}\ \text{m}$

Exercise:

Problem:

Find the distance between two slits that produces the first minimum for 410-nm violet light at an angle of 45.0^{o}.

Exercise:

Problem:

Calculate the wavelength of light that has its third minimum at an angle of 30.0^{o} when falling on double slits separated by $3.00\ \mu\text{m}$. Explicitly, show how you follow the steps in [Problem-Solving Strategies for Wave Optics].

Solution:

600 nm

Exercise:

Problem:

What is the wavelength of light falling on double slits separated by $2.00\ \mu\text{m}$ if the third-order maximum is at an angle of 60.0^{o}?

Exercise:

Problem:

At what angle is the fourth-order maximum for the situation in [link]?

Solution:

2.06°

Exercise:

Problem:

What is the highest-order maximum for 400-nm light falling on double slits separated by $25.0\ \mu m$?

Exercise:

Problem:

Find the largest wavelength of light falling on double slits separated by $1.20\ \mu m$ for which there is a first-order maximum. Is this in the visible part of the spectrum?

Solution:

1200 nm (not visible)

Exercise:

Problem:

What is the smallest separation between two slits that will produce a second-order maximum for 720-nm red light?

Exercise:

Problem:

(a) What is the smallest separation between two slits that will produce a second-order maximum for any visible light? (b) For all visible light?

Solution:

(a) 760 nm

(b) 1520 nm

Exercise:

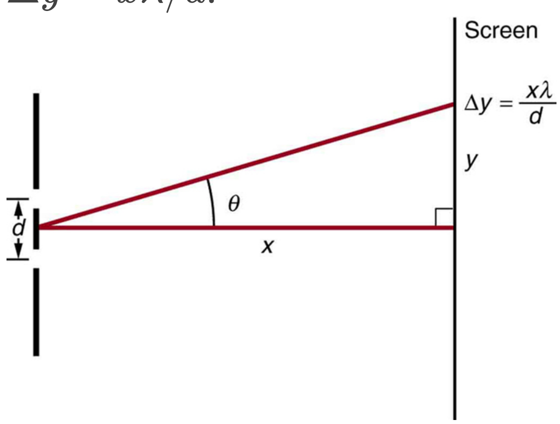
Problem:

(a) If the first-order maximum for pure-wavelength light falling on a double slit is at an angle of 10.0^{o}, at what angle is the second-order maximum? (b) What is the angle of the first minimum? (c) What is the highest-order maximum possible here?

Exercise:

Problem:

[link] shows a double slit located a distance x from a screen, with the distance from the center of the screen given by y. When the distance d between the slits is relatively large, there will be numerous bright spots, called fringes. Show that, for small angles (where $\sin \theta \approx \theta$, with θ in radians), the distance between fringes is given by $\Delta y = x\lambda/d$.



The distance between adjacent fringes is $\Delta y = x\lambda/d$, assuming the

slit separation d is large
compared with λ.

Solution:

For small angles $\sin\,\theta - \tan\,\theta \approx \theta$ (in radians).

For two adjacent fringes we have,

Equation:

$$d\ \ \sin\,\theta_{\text{m}} = m\lambda$$

and

Equation:

$$d\ \ \sin\,\theta_{\text{m}+1} = (m+1)\lambda$$

Subtracting these equations gives

Equation:

$$\begin{gathered}d(\sin\,\theta_{\text{m}+1} - \sin\,\theta_{\text{m}}) = [(m+1) - m]\lambda \\ d(\theta_{\text{m}+1} - \theta_{\text{m}}) = \lambda \\ \tan\,\theta_{\text{m}} = \frac{y_{\text{m}}}{x} \approx \theta_{\text{m}} \Rightarrow d\ \ \frac{y_{\text{m}+1}}{x} - \frac{y_{\text{m}}}{x}\ \ = \lambda \\ d\frac{\Delta y}{x} = \lambda \Rightarrow \Delta y = \frac{x\lambda}{d}\end{gathered}$$

Exercise:

Problem:

Using the result of the problem above, calculate the distance between fringes for 633-nm light falling on double slits separated by 0.0800 mm, located 3.00 m from a screen as in [link].

Exercise:

Problem:

Using the result of the problem two problems prior, find the wavelength of light that produces fringes 7.50 mm apart on a screen 2.00 m from double slits separated by 0.120 mm (see [link]).

Solution:

450 nm

Glossary

coherent
: waves are in phase or have a definite phase relationship

constructive interference for a double slit
: the path length difference must be an integral multiple of the wavelength

destructive interference for a double slit
: the path length difference must be a half-integral multiple of the wavelength

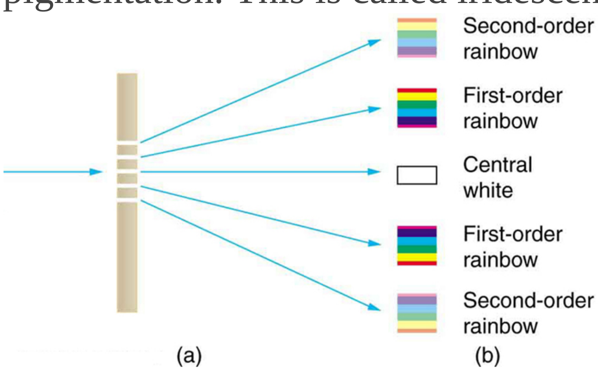
incoherent
: waves have random phase relationships

order
: the integer m used in the equations for constructive and destructive interference for a double slit

Multiple Slit Diffraction

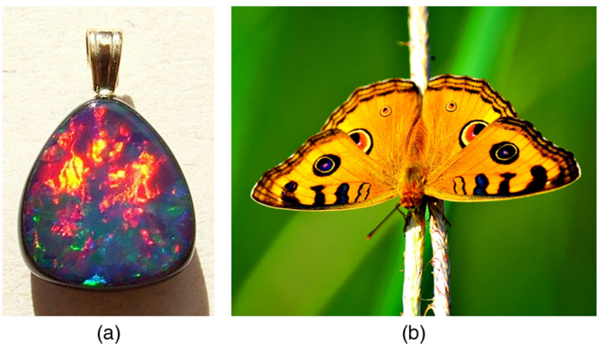
- Discuss the pattern obtained from diffraction grating.
- Explain diffraction grating effects.

An interesting thing happens if you pass light through a large number of evenly spaced parallel slits, called a **diffraction grating**. An interference pattern is created that is very similar to the one formed by a double slit (see [link]). A diffraction grating can be manufactured by scratching glass with a sharp tool in a number of precisely positioned parallel lines, with the untouched regions acting like slits. These can be photographically mass produced rather cheaply. Diffraction gratings work both for transmission of light, as in [link], and for reflection of light, as on butterfly wings and the Australian opal in [link] or the CD pictured in the opening photograph of this chapter, [link]. In addition to their use as novelty items, diffraction gratings are commonly used for spectroscopic dispersion and analysis of light. What makes them particularly useful is the fact that they form a sharper pattern than double slits do. That is, their bright regions are narrower and brighter, while their dark regions are darker. [link] shows idealized graphs demonstrating the sharper pattern. Natural diffraction gratings occur in the feathers of certain birds. Tiny, finger-like structures in regular patterns act as reflection gratings, producing constructive interference that gives the feathers colors not solely due to their pigmentation. This is called iridescence.

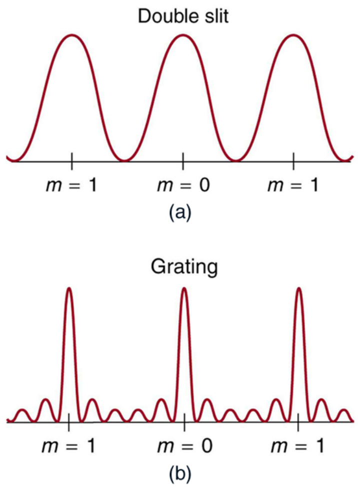


A diffraction grating is a large number of evenly spaced parallel slits. (a) Light passing through is diffracted in a pattern similar to a double slit, with

bright regions at various angles. (b) The pattern obtained for white light incident on a grating. The central maximum is white, and the higher-order maxima disperse white light into a rainbow of colors.



(a) This Australian opal and (b) the butterfly wings have rows of reflectors that act like reflection gratings, reflecting different colors at different angles. (credits: (a) Opals-On-Black.com, via Flickr (b) whologwhy, Flickr)



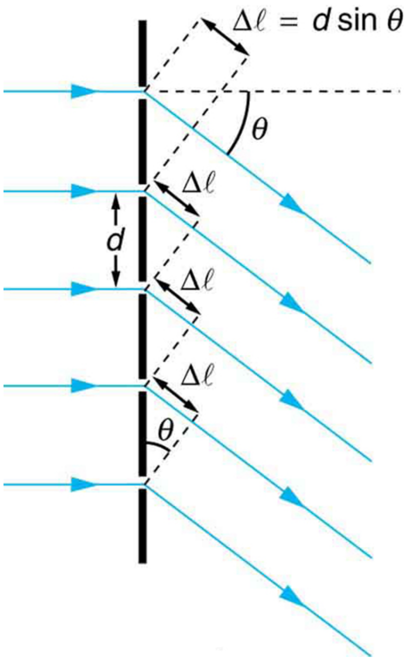
Idealized graphs of the intensity of light passing through a double slit (a) and a diffraction grating (b) for monochromatic light. Maxima can be produced at the same angles, but those for the diffraction grating are narrower and hence sharper. The maxima become narrower and the regions between darker as the number of slits is increased.

The analysis of a diffraction grating is very similar to that for a double slit (see [link]). As we know from our discussion of double slits in Young's Double Slit Experiment, light is diffracted by each slit and spreads out after passing through. Rays traveling in the same direction (at an angle θ relative to the incident direction) are shown in the figure. Each of these rays travels a different distance to a common point on a screen far away. The rays start in phase, and they can be in or out of phase when they reach a screen, depending on the difference in the path lengths traveled. As seen in the figure, each ray travels a distance $d \sin \theta$ different from that of its neighbor, where d is the distance between slits. If this distance equals an integral number of wavelengths, the rays all arrive in phase, and constructive interference (a maximum) is obtained. Thus, the condition necessary to obtain **constructive interference for a diffraction grating** is

Equation:

$$d \ \sin \theta = m\lambda, \ \text{for} \ m = 0, 1, –1, 2, –2, \ldots \ (\text{constructive}),$$

where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum. Note that this is exactly the same equation as for double slits separated by d. However, the slits are usually closer in diffraction gratings than in double slits, producing fewer maxima at larger angles.



Diffraction grating showing light rays from each slit traveling in the same direction. Each ray travels a different distance to reach a common point on a screen (not shown). Each ray travels a distance $d \sin \theta$ different from that of its neighbor.

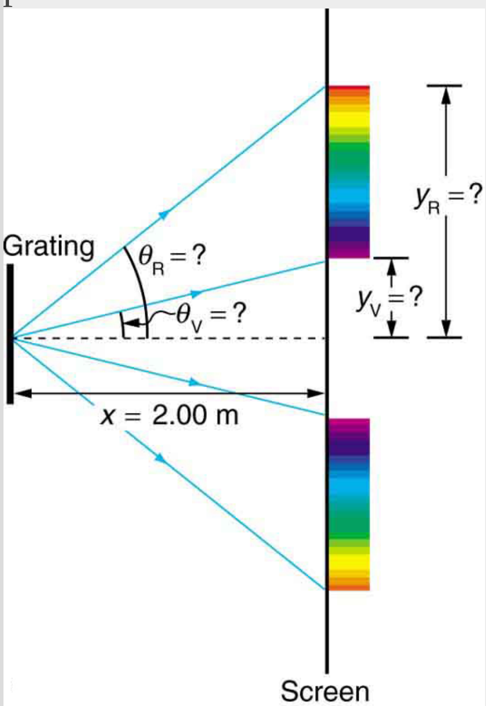
Where are diffraction gratings used? Diffraction gratings are key components of monochromators used, for example, in optical imaging of particular wavelengths from biological or medical samples. A diffraction grating can be chosen to specifically analyze a wavelength emitted by molecules in diseased cells in a biopsy sample or to help excite strategic molecules in the sample with a selected frequency of light. Another vital use is in optical fiber technologies where fibers are designed to provide optimum performance at specific wavelengths. A range of diffraction gratings are available for selecting specific wavelengths for such use.

Note:

Take-Home Experiment: Rainbows on a CD

The spacing d of the grooves in a CD or DVD can be well determined by using a laser and the equation $d \sin \theta = m\lambda$, for $m = 0, 1, -1, 2, -2, \ldots$. However, we can still make a good estimate of this spacing by using white light and the rainbow of colors that comes from the interference. Reflect sunlight from a CD onto a wall and use your best judgment of the location of a strongly diffracted color to find the separation d.

Example:
Calculating Typical Diffraction Grating Effects
Diffraction gratings with 10,000 lines per centimeter are readily available. Suppose you have one, and you send a beam of white light through it to a screen 2.00 m away. (a) Find the angles for the first-order diffraction of the shortest and longest wavelengths of visible light (380 and 760 nm). (b) What is the distance between the ends of the rainbow of visible light produced on the screen for first-order interference? (See [link].)



The diffraction grating considered in this example produces a rainbow of colors on a screen a distance $x = 2.00\ \text{m}$ from the grating. The distances along the screen are measured perpendicular to the x-direction. In other words, the rainbow pattern extends out of the page.

Strategy

The angles can be found using the equation

Equation:

$$d\ \sin\theta = m\lambda\ (\text{for } m = 0, 1, –1, 2, –2,\ \ldots)$$

once a value for the slit spacing d has been determined. Since there are 10,000 lines per centimeter, each line is separated by $1/10{,}000$ of a centimeter. Once the angles are found, the distances along the screen can be found using simple trigonometry.

Solution for (a)

The distance between slits is $d = (1\text{ cm})/10{,}000 = 1.00 \times 10^{-4}$ cm or 1.00×10^{-6} m. Let us call the two angles θ_{V} for violet (380 nm) and θ_{R} for red (760 nm). Solving the equation d sin $\theta_{\text{V}} = m\lambda$ for sin θ_{V},

Equation:

$$\sin\theta_{\text{V}} = \frac{m\lambda_{\text{V}}}{d},$$

where $m = 1$ for first order and $\lambda_{\text{V}} = 380\text{ nm} = 3.80 \times 10^{-7}$ m. Substituting these values gives

Equation:

$$\sin\theta_{\text{V}} = \frac{3.80 \times 10^{-7}\text{ m}}{1.00 \times 10^{-6}\text{ m}} = 0.380.$$

Thus the angle θ_{V} is

Equation:

$$\theta_{\text{V}} = \sin^{-1} 0.380 = 22.33^{\text{o}}.$$

Similarly,

Equation:

$$\sin\theta_{\text{R}} = \frac{7.60 \times 10^{-7}\text{ m}}{1.00 \times 10^{-6}\text{ m}}.$$

Thus the angle θ_{R} is
Equation:

$$\theta_{\text{R}} = \sin^{-1} 0.760 = 49.46^{\text{o}}.$$

Notice that in both equations, we reported the results of these intermediate calculations to four significant figures to use with the calculation in part (b).

Solution for (b)

The distances on the screen are labeled y_{V} and y_{R} in [link]. Noting that $\tan\,\theta = y/x$, we can solve for y_{V} and y_{R}. That is,
Equation:

$$y_{\text{V}} = x\,\tan\,\theta_{\text{V}} = (2.00\text{ m})(\tan\,22.33^{\text{o}}) = 0.815\text{ m}$$

and
Equation:

$$y_{\text{R}} = x\,\tan\,\theta_{\text{R}} = (2.00\text{ m})(\tan\,49.46^{\text{o}}) = 2.338\text{ m}.$$

The distance between them is therefore
Equation:

$$y_{\text{R}} - y_{\text{V}} = 1.52\text{ m}.$$

Discussion

The large distance between the red and violet ends of the rainbow produced from the white light indicates the potential this diffraction grating has as a spectroscopic tool. The more it can spread out the wavelengths (greater dispersion), the more detail can be seen in a spectrum. This depends on the quality of the diffraction grating—it must be very precisely made in addition to having closely spaced lines.

Section Summary

- A diffraction grating is a large collection of evenly spaced parallel slits that produces an interference pattern similar to but sharper than that of a double slit.
- There is constructive interference for a diffraction grating when $d \ \sin \theta = m\lambda$ (for $m = 0, 1, –1, 2, –2, \ \ldots$), where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum.

Conceptual Questions

Exercise:

Problem:

What is the advantage of a diffraction grating over a double slit in dispersing light into a spectrum?

Exercise:

Problem:

What are the advantages of a diffraction grating over a prism in dispersing light for spectral analysis?

Exercise:

Problem:

Can the lines in a diffraction grating be too close together to be useful as a spectroscopic tool for visible light? If so, what type of EM radiation would the grating be suitable for? Explain.

Exercise:

Problem:

If a beam of white light passes through a diffraction grating with vertical lines, the light is dispersed into rainbow colors on the right and left. If a glass prism disperses white light to the right into a rainbow, how does the sequence of colors compare with that produced on the right by a diffraction grating?

Exercise:

Problem:

Suppose pure-wavelength light falls on a diffraction grating. What happens to the interference pattern if the same light falls on a grating that has more lines per centimeter? What happens to the interference pattern if a longer-wavelength light falls on the same grating? Explain how these two effects are consistent in terms of the relationship of wavelength to the distance between slits.

Exercise:

Problem:

Suppose a feather appears green but has no green pigment. Explain in terms of diffraction.

Exercise:

Problem:

It is possible that there is no minimum in the interference pattern of a single slit. Explain why. Is the same true of double slits and diffraction gratings?

Problems & Exercises

Exercise:

Problem:

A diffraction grating has 2000 lines per centimeter. At what angle will the first-order maximum be for 520-nm-wavelength green light?

Solution:

5.97°

Exercise:

Problem:

Find the angle for the third-order maximum for 580-nm-wavelength yellow light falling on a diffraction grating having 1500 lines per centimeter.

Exercise:

Problem:

How many lines per centimeter are there on a diffraction grating that gives a first-order maximum for 470-nm blue light at an angle of 25.0° ?

Solution:

8.99×10^{3}

Exercise:

Problem:

What is the distance between lines on a diffraction grating that produces a second-order maximum for 760-nm red light at an angle of 60.0°?

Exercise:

Problem:

Calculate the wavelength of light that has its second-order maximum at 45.0° when falling on a diffraction grating that has 5000 lines per centimeter.

Solution:

707 nm

Exercise:

Problem:

An electric current through hydrogen gas produces several distinct wavelengths of visible light. What are the wavelengths of the hydrogen spectrum, if they form first-order maxima at angles of 24.2°, 25.7°, 29.1°, and 41.0° when projected on a diffraction grating having 10,000 lines per centimeter? Explicitly show how you follow the steps in Problem-Solving Strategies for Wave Optics

Exercise:

Problem:

(a) What do the four angles in the above problem become if a 5000-line-per-centimeter diffraction grating is used? (b) Using this grating, what would the angles be for the second-order maxima? (c) Discuss the relationship between integral reductions in lines per centimeter and the new angles of various order maxima.

Solution:

$$(a)\ \ 11.8^{\circ},\ 12.5^{\circ},\ 14.1^{\circ},\ 19.2^{\circ}$$

$$(b)\ \ 24.2^{\circ},\ 25.7^{\circ},\ 29.1^{\circ},\ 41.0^{\circ}$$

(c) Decreasing the number of lines per centimeter by a factor of x means that the angle for the x-order maximum is the same as the original angle for the first- order maximum.

Exercise:

Problem:

What is the maximum number of lines per centimeter a diffraction grating can have and produce a complete first-order spectrum for visible light?

Exercise:

Problem:

The yellow light from a sodium vapor lamp *seems* to be of pure wavelength, but it produces two first-order maxima at 36.093° and 36.129° when projected on a 10,000 line per centimeter diffraction grating. What are the two wavelengths to an accuracy of 0.1 nm?

Solution:

589.1 nm and 589.6 nm

Exercise:

Problem:

What is the spacing between structures in a feather that acts as a reflection grating, given that they produce a first-order maximum for 525-nm light at a 30.0° angle?

Exercise:

Problem:

Structures on a bird feather act like a reflection grating having 8000 lines per centimeter. What is the angle of the first-order maximum for 600-nm light?

Solution:

28.7°

Exercise:

Problem:

An opal such as that shown in [link] acts like a reflection grating with rows separated by about $8\ \mu m$. If the opal is illuminated normally, (a) at what angle will red light be seen and (b) at what angle will blue light be seen?

Exercise:

Problem:

At what angle does a diffraction grating produces a second-order maximum for light having a first-order maximum at 20.0^o?

Solution:

43.2^o

Exercise:

Problem:

Show that a diffraction grating cannot produce a second-order maximum for a given wavelength of light unless the first-order maximum is at an angle less than 30.0^o.

Exercise:

Problem:

If a diffraction grating produces a first-order maximum for the shortest wavelength of visible light at 30.0^o, at what angle will the first-order maximum be for the longest wavelength of visible light?

Solution:

90.0^o

Exercise:

Problem:

(a) Find the maximum number of lines per centimeter a diffraction grating can have and produce a maximum for the smallest wavelength of visible light. (b) Would such a grating be useful for ultraviolet spectra? (c) For infrared spectra?

Exercise:

Problem:

(a) Show that a 30,000-line-per-centimeter grating will not produce a maximum for visible light. (b) What is the longest wavelength for which it does produce a first-order maximum? (c) What is the greatest number of lines per centimeter a diffraction grating can have and produce a complete second-order spectrum for visible light?

Solution:

(a) The longest wavelength is 333.3 nm, which is not visible.

(b) 333 nm (UV)

(c) 6.58×10^3 cm

Exercise:

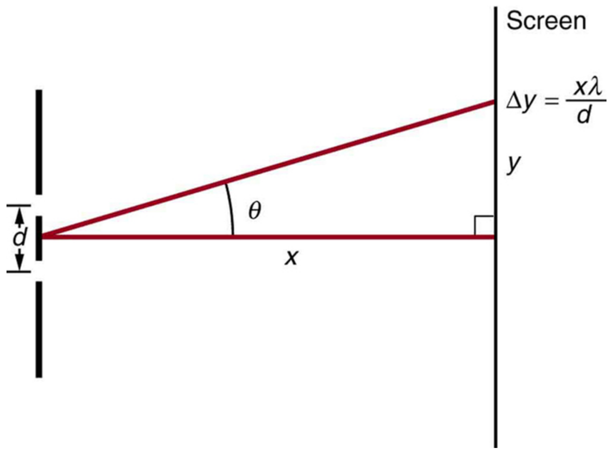
Problem:

A He–Ne laser beam is reflected from the surface of a CD onto a wall. The brightest spot is the reflected beam at an angle equal to the angle of incidence. However, fringes are also observed. If the wall is 1.50 m from the CD, and the first fringe is 0.600 m from the central maximum, what is the spacing of grooves on the CD?

Exercise:

Problem:

The analysis shown in the figure below also applies to diffraction gratings with lines separated by a distance d. What is the distance between fringes produced by a diffraction grating having 125 lines per centimeter for 600-nm light, if the screen is 1.50 m away?



The distance between adjacent fringes is $\Delta y = x\lambda/d$, assuming the slit separation d is large compared with λ.

Solution:

1.13×10^{-2} m

Exercise:

Problem: Unreasonable Results

Red light of wavelength of 700 nm falls on a double slit separated by 400 nm. (a) At what angle is the first-order maximum in the diffraction pattern? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: Unreasonable Results

(a) What visible wavelength has its fourth-order maximum at an angle of 25.0° when projected on a 25,000-line-per-centimeter diffraction

grating? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) 42.3 nm

(b) Not a visible wavelength

The number of slits in this diffraction grating is too large. Etching in integrated circuits can be done to a resolution of 50 nm, so slit separations of 400 nm are at the limit of what we can do today. This line spacing is too small to produce diffraction of light.

Exercise:

Problem: Construct Your Own Problem

Consider a spectrometer based on a diffraction grating. Construct a problem in which you calculate the distance between two wavelengths of electromagnetic radiation in your spectrometer. Among the things to be considered are the wavelengths you wish to be able to distinguish, the number of lines per meter on the diffraction grating, and the distance from the grating to the screen or detector. Discuss the practicality of the device in terms of being able to discern between wavelengths of interest.

Glossary

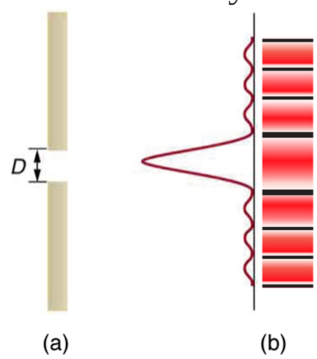
constructive interference for a diffraction grating
: occurs when the condition $d\ \sin\theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ \ldots$) is satisfied, where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum

diffraction grating
: a large number of evenly spaced parallel slits

Single Slit Diffraction

- Discuss the single slit diffraction pattern.

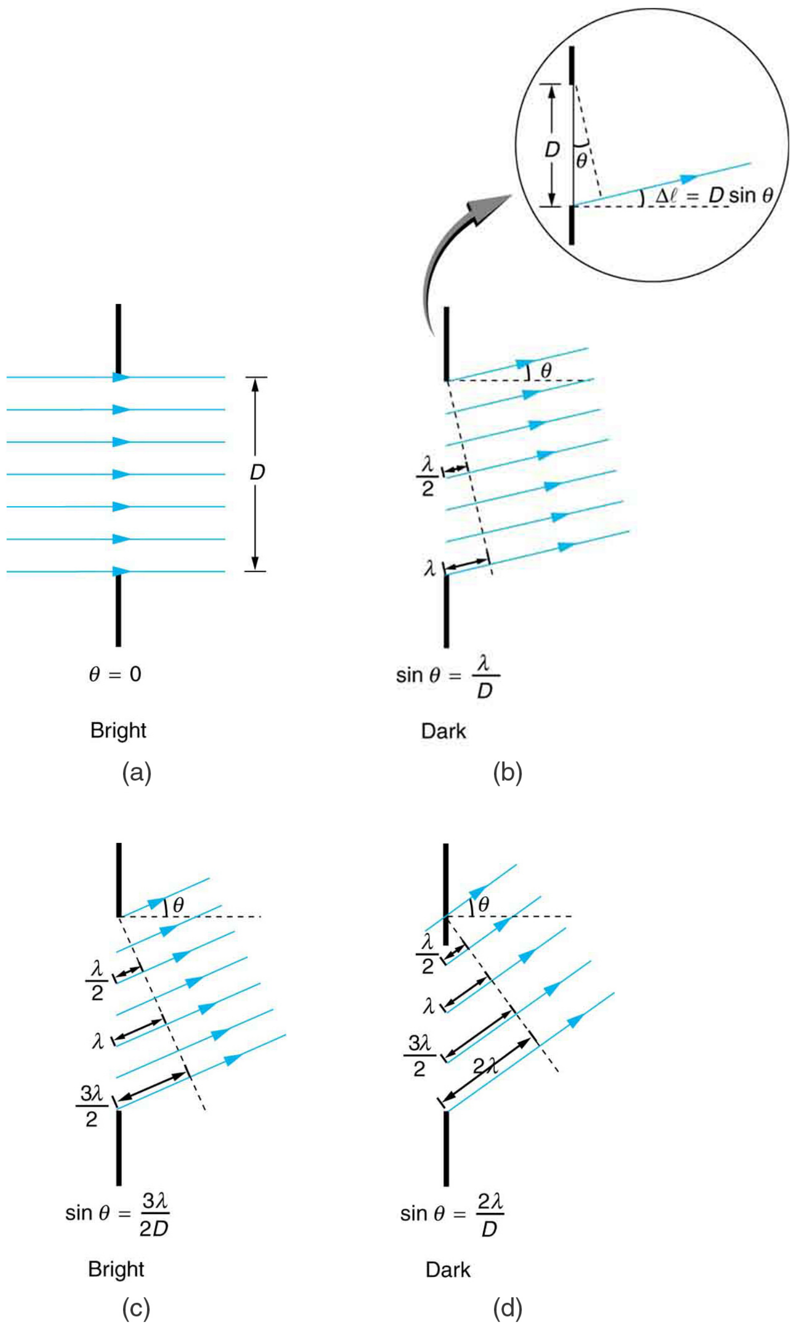
Light passing through a single slit forms a diffraction pattern somewhat different from those formed by double slits or diffraction gratings. [link] shows a single slit diffraction pattern. Note that the central maximum is larger than those on either side, and that the intensity decreases rapidly on either side. In contrast, a diffraction grating produces evenly spaced lines that dim slowly on either side of center.



(a) Single slit diffraction pattern. Monochromatic light passing through a single slit has a central maximum and many smaller and dimmer maxima on either side. The central maximum is six times higher than shown. (b) The drawing shows

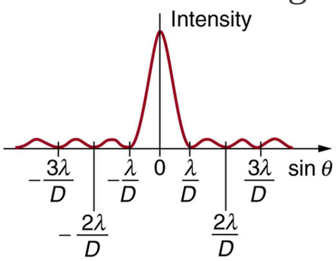
the bright central maximum and dimmer and thinner maxima on either side.

The analysis of single slit diffraction is illustrated in [link]. Here we consider light coming from different parts of the *same* slit. According to Huygens's principle, every part of the wavefront in the slit emits wavelets. These are like rays that start out in phase and head in all directions. (Each ray is perpendicular to the wavefront of a wavelet.) Assuming the screen is very far away compared with the size of the slit, rays heading toward a common destination are nearly parallel. When they travel straight ahead, as in [link](a), they remain in phase, and a central maximum is obtained. However, when rays travel at an angle θ relative to the original direction of the beam, each travels a different distance to a common location, and they can arrive in or out of phase. In [link](b), the ray from the bottom travels a distance of one wavelength λ farther than the ray from the top. Thus a ray from the center travels a distance λ farther than the one on the left, arrives out of phase, and interferes destructively. A ray from slightly above the center and one from slightly above the bottom will also cancel one another. In fact, each ray from the slit will have another to interfere destructively, and a minimum in intensity will occur at this angle. There will be another minimum at the same angle to the right of the incident direction of the light.



Light passing through a single slit is diffracted in all directions and may interfere constructively or destructively, depending on the angle. The difference in path length for rays from either side of the slit is seen to be $D \quad \theta$.

At the larger angle shown in [link](c), the path lengths differ by λ for rays from the top and bottom of the slit. One ray travels a distance λ different from the ray from the bottom and arrives in phase, interfering constructively. Two rays, each from slightly above those two, will also add constructively. Most rays from the slit will have another to interfere with constructively, and a maximum in intensity will occur at this angle. However, all rays do not interfere constructively for this situation, and so the maximum is not as intense as the central maximum. Finally, in [link](d), the angle shown is large enough to produce a second minimum. As seen in the figure, the difference in path length for rays from either side of the slit is $D \quad \theta$, and we see that a destructive minimum is obtained when this distance is an integral multiple of the wavelength.



A graph of single slit diffraction intensity showing the central maximum to be wider and much more intense than those to the sides. In fact the central maximum is six times higher than shown here.

Thus, to obtain **destructive interference for a single slit**,

Equation:

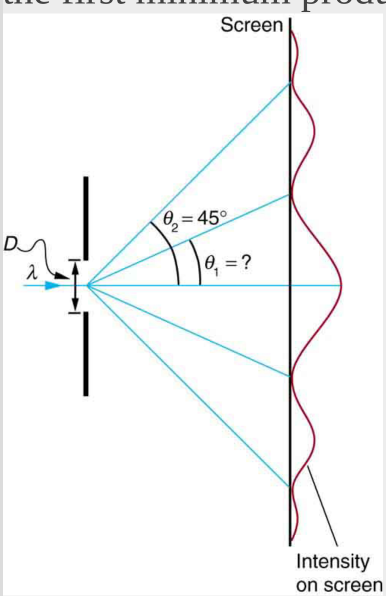
$$D \qquad \theta \qquad m\lambda \qquad m$$

where D is the slit width, λ is the light’s wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum. [link] shows a graph of intensity for single slit interference, and it is apparent that the maxima on either side of the central maximum are much less intense and not as wide. This is consistent with the illustration in [link] (b).

Example:

Calculating Single Slit Diffraction

Visible light of wavelength 550 nm falls on a single slit and produces its second diffraction minimum at an angle of º relative to the incident direction of the light. (a) What is the width of the slit? (b) At what angle is the first minimum produced?



A graph of the

single slit diffraction pattern is analyzed in this example.

Strategy

From the given information, and assuming the screen is far away from the slit, we can use the equation $D \quad \theta \quad$ λ first to find D, and again to find the angle for the first minimum θ .

Solution for (a)

We are given that λ , m , and θ º. Solving the equation $D \quad \theta \quad$ λ for D and substituting known values gives

Equation:

$$D \quad \frac{m\lambda}{\theta} \quad \frac{\quad}{\quad ^\circ}$$

$$\frac{\quad}{\quad}$$

Solution for (b)

Solving the equation $D \quad \theta \quad m\lambda$ for $\quad \theta$ and substituting the known values gives

Equation:

$$\theta \quad \frac{m\lambda}{D} \quad \frac{\quad}{\quad}$$

Thus the angle θ is

Equation:

$$\theta \quad ^\circ$$

Discussion

We see that the slit is narrow (it is only a few times greater than the wavelength of light). This is consistent with the fact that light must interact with an object comparable in size to its wavelength in order to exhibit

significant wave effects such as this single slit diffraction pattern. We also see that the central maximum extends º on either side of the original beam, for a width of about º . The angle between the first and second minima is only about º º º . Thus the second maximum is only about half as wide as the central maximum.

Section Summary

- A single slit produces an interference pattern characterized by a broad central maximum with narrower and dimmer maxima to the sides.
- There is destructive interference for a single slit when D θ $m\lambda$ m , where D is the slit width, λ is the light's wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum. Note that there is no m minimum.

Conceptual Questions

Exercise:

Problem:

As the width of the slit producing a single-slit diffraction pattern is reduced, how will the diffraction pattern produced change?

Problems & Exercises

Exercise:

Problem:

(a) At what angle is the first minimum for 550-nm light falling on a single slit of width μ ? (b) Will there be a second minimum?

Solution:

(a) º

(b) No

Exercise:

Problem:

(a) Calculate the angle at which a μ -wide slit produces its first minimum for 410-nm violet light. (b) Where is the first minimum for 700-nm red light?

Exercise:

Problem:

(a) How wide is a single slit that produces its first minimum for 633-nm light at an angle of º? (b) At what angle will the second minimum be?

Solution:

(a)

(b) º

Exercise:

Problem:

(a) What is the width of a single slit that produces its first minimum at º for 600-nm light? (b) Find the wavelength of light that has its first minimum at º.

Exercise:

Problem:

Find the wavelength of light that has its third minimum at an angle of º when it falls on a single slit of width μ .

Solution:

750 nm

Exercise:

Problem:

Calculate the wavelength of light that produces its first minimum at an angle of º when falling on a single slit of width μ .

Exercise:

Problem:

(a) Sodium vapor light averaging 589 nm in wavelength falls on a single slit of width μ . At what angle does it produces its second minimum? (b) What is the highest-order minimum produced?

Solution:

(a) º

(b) 12

Exercise:

Problem:

(a) Find the angle of the third diffraction minimum for 633-nm light falling on a slit of width μ . (b) What slit width would place this minimum at º? Explicitly show how you follow the steps in Problem-Solving Strategies for Wave Optics

Exercise:

Problem:

(a) Find the angle between the first minima for the two sodium vapor lines, which have wavelengths of 589.1 and 589.6 nm, when they fall upon a single slit of width μ . (b) What is the distance between these minima if the diffraction pattern falls on a screen 1.00 m from the slit? (c) Discuss the ease or difficulty of measuring such a distance.

Solution:

(a) º

(b) 0.262 mm

(c) This distance is not easily measured by human eye, but under a microscope or magnifying glass it is quite easily measurable.

Exercise:

Problem:

(a) What is the minimum width of a single slit (in multiples of λ) that will produce a first minimum for a wavelength λ? (b) What is its minimum width if it produces 50 minima? (c) 1000 minima?

Exercise:

Problem:

(a) If a single slit produces a first minimum at º, at what angle is the second-order minimum? (b) What is the angle of the third-order minimum? (c) Is there a fourth-order minimum? (d) Use your answers to illustrate how the angular width of the central maximum is about twice the angular width of the next maximum (which is the angle between the first and second minima).

Solution:

(a) º

(b) º

(c) No

(d) θ º º θ θ º º º. Thus, º º º.

Exercise:

Problem:

A double slit produces a diffraction pattern that is a combination of single and double slit interference. Find the ratio of the width of the slits to the separation between them, if the first minimum of the single slit pattern falls on the fifth maximum of the double slit pattern. (This will greatly reduce the intensity of the fifth maximum.)

Exercise:

Problem: Integrated Concepts

A water break at the entrance to a harbor consists of a rock barrier with a 50.0-m-wide opening. Ocean waves of 20.0-m wavelength approach the opening straight on. At what angle to the incident direction are the boats inside the harbor most protected against wave action?

Solution:

º and º

Exercise:

Problem: Integrated Concepts

An aircraft maintenance technician walks past a tall hangar door that acts like a single slit for sound entering the hangar. Outside the door, on a line perpendicular to the opening in the door, a jet engine makes a 600-Hz sound. At what angle with the door will the technician observe the first minimum in sound intensity if the vertical opening is 0.800 m wide and the speed of sound is 340 m/s?

Glossary

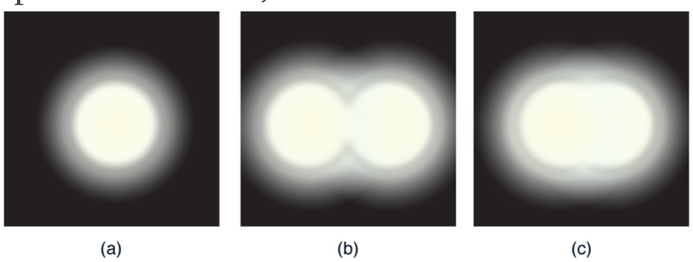
destructive interference for a single slit
: occurs when D θ λ m , where D is the slit width, λ is the light's wavelength, θ is the angle relative to

the original direction of the light, and m is the order of the minimum

Limits of Resolution: The Rayleigh Criterion

- Discuss the Rayleigh criterion.

Light diffracts as it moves through space, bending around obstacles, interfering constructively and destructively. While this can be used as a spectroscopic tool—a diffraction grating disperses light according to wavelength, for example, and is used to produce spectra—diffraction also limits the detail we can obtain in images. [link](a) shows the effect of passing light through a small circular aperture. Instead of a bright spot with sharp edges, a spot with a fuzzy edge surrounded by circles of light is obtained. This pattern is caused by diffraction similar to that produced by a single slit. Light from different parts of the circular aperture interferes constructively and destructively. The effect is most noticeable when the aperture is small, but the effect is there for large apertures, too.



(a) Monochromatic light passed through a small circular aperture produces this diffraction pattern. (b) Two point light sources that are close to one another produce overlapping images because of diffraction. (c) If they are closer together, they cannot be resolved or distinguished.

How does diffraction affect the detail that can be observed when light passes through an aperture? [link](b) shows the diffraction pattern produced by two point light sources that are close to one another. The pattern is similar to that for a single point source, and it is just barely possible to tell that there are two light sources rather than one. If they were closer together,

as in [link](c), we could not distinguish them, thus limiting the detail or resolution we can obtain. This limit is an inescapable consequence of the wave nature of light.

There are many situations in which diffraction limits the resolution. The acuity of our vision is limited because light passes through the pupil, the circular aperture of our eye. Be aware that the diffraction-like spreading of light is due to the limited diameter of a light beam, not the interaction with an aperture. Thus light passing through a lens with a diameter D shows this effect and spreads, blurring the image, just as light passing through an aperture of diameter D does. So diffraction limits the resolution of any system having a lens or mirror. Telescopes are also limited by diffraction, because of the finite diameter D of their primary mirror.

Note:
Take-Home Experiment: Resolution of the Eye
Draw two lines on a white sheet of paper (several mm apart). How far away can you be and still distinguish the two lines? What does this tell you about the size of the eye's pupil? Can you be quantitative? (The size of an adult's pupil is discussed in Physics of the Eye.)

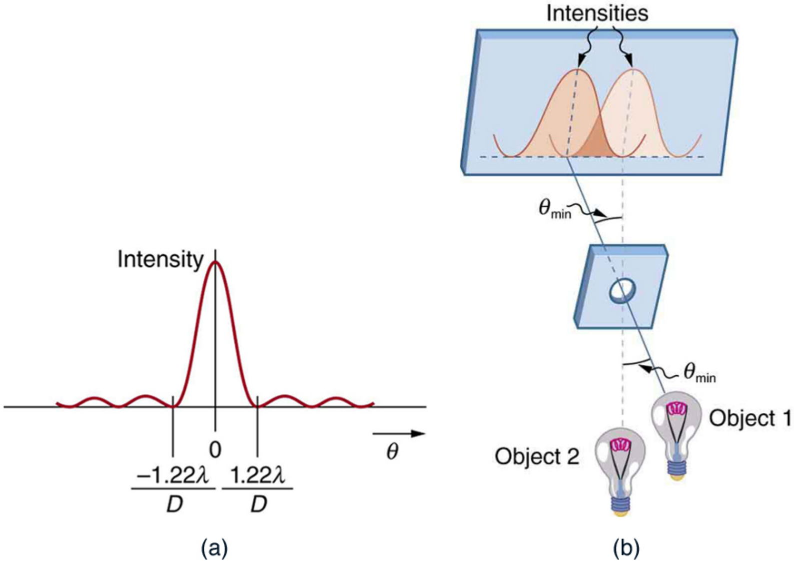
Just what is the limit? To answer that question, consider the diffraction pattern for a circular aperture, which has a central maximum that is wider and brighter than the maxima surrounding it (similar to a slit) [see [link](a)]. It can be shown that, for a circular aperture of diameter D, the first minimum in the diffraction pattern occurs at $\theta = 1.22\ \lambda/D$ (providing the aperture is large compared with the wavelength of light, which is the case for most optical instruments). The accepted criterion for determining the diffraction limit to resolution based on this angle was developed by Lord Rayleigh in the 19th century. The **Rayleigh criterion** for the diffraction limit to resolution states that *two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other*. See [link](b). The first minimum is at an

angle of $\theta = 1.22\ \lambda/D$, so that two point objects are just resolvable if they are separated by the angle

Equation:

$$\theta = 1.22\frac{\lambda}{D},$$

where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc., with which the two objects are observed. In this expression, θ has units of radians.



(a) Graph of intensity of the diffraction pattern for a circular aperture. Note that, similar to a single slit, the central maximum is wider and brighter than those to the sides. (b) Two point objects produce overlapping diffraction patterns. Shown here is the Rayleigh criterion for being just resolvable. The central maximum of one pattern lies on the first minimum of the other.

Note:
Connections: Limits to Knowledge
All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Even the small wavelength of light prohibits exact precision. When extremely small wavelength probes as with an electron microscope are used, the system is disturbed, still limiting our knowledge, much as making an electrical measurement alters a circuit. Heisenberg's uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in quantum mechanics.

Example:
Calculating Diffraction Limits of the Hubble Space Telescope
The primary mirror of the orbiting Hubble Space Telescope has a diameter of 2.40 m. Being in orbit, this telescope avoids the degrading effects of atmospheric distortion on its resolution. (a) What is the angle between two just-resolvable point light sources (perhaps two stars)? Assume an average light wavelength of 550 nm. (b) If these two stars are at the 2 million light year distance of the Andromeda galaxy, how close together can they be and still be resolved? (A light year, or ly, is the distance light travels in 1 year.)
Strategy
The Rayleigh criterion stated in the equation $\theta = 1.22\frac{\lambda}{D}$ gives the smallest possible angle θ between point sources, or the best obtainable resolution. Once this angle is found, the distance between stars can be calculated, since we are given how far away they are.
Solution for (a)
The Rayleigh criterion for the minimum resolvable angle is
Equation:

$$\theta = 1.22\frac{\lambda}{D}.$$

Entering known values gives
Equation:

$$\theta = 1.22\frac{550\times10^{-9}\text{ m}}{2.40\text{ m}} = 2.80\times10^{-7}\text{ rad}.$$

Solution for (b)

The distance s between two objects a distance r away and separated by an angle θ is $s = r\theta$.

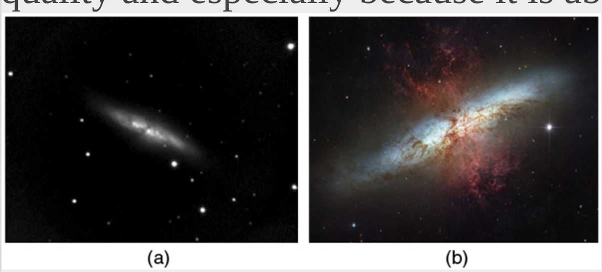
Substituting known values gives

Equation:

$$\begin{aligned} s &= (2.0\times10^{6}\text{ ly})(2.80\times10^{-7}\text{ rad}) \\ &= 0.56\text{ ly}. \end{aligned}$$

Discussion

The angle found in part (a) is extraordinarily small (less than 1/50,000 of a degree), because the primary mirror is so large compared with the wavelength of light. As noticed, diffraction effects are most noticeable when light interacts with objects having sizes on the order of the wavelength of light. However, the effect is still there, and there is a diffraction limit to what is observable. The actual resolution of the Hubble Telescope is not quite as good as that found here. As with all instruments, there are other effects, such as non-uniformities in mirrors or aberrations in lenses that further limit resolution. However, [link] gives an indication of the extent of the detail observable with the Hubble because of its size and quality and especially because it is above the Earth's atmosphere.



These two photographs of the M82 galaxy give an idea of the observable detail using the Hubble Space Telescope compared with that using a ground-based telescope. (a) On

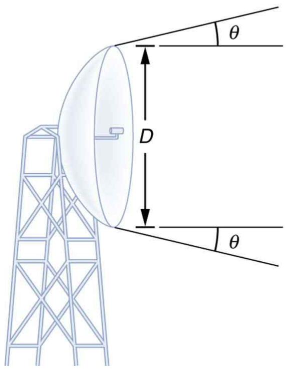
the left is a ground-based image. (credit: Ricnun, Wikimedia Commons) (b) The photo on the right was captured by Hubble. (credit: NASA, ESA, and the Hubble Heritage Team (STScI/AURA))

The answer in part (b) indicates that two stars separated by about half a light year can be resolved. The average distance between stars in a galaxy is on the order of 5 light years in the outer parts and about 1 light year near the galactic center. Therefore, the Hubble can resolve most of the individual stars in Andromeda galaxy, even though it lies at such a huge distance that its light takes 2 million years for its light to reach us. [link] shows another mirror used to observe radio waves from outer space.

A 305-m-diameter natural bowl at Arecibo in Puerto Rico is lined with reflective material, making it into a radio telescope. It is the largest curved focusing dish in the world. Although D for Arecibo is much larger than for the

Hubble Telescope, it detects much longer wavelength radiation and its diffraction limit is significantly poorer than Hubble's. Arecibo is still very useful, because important information is carried by radio waves that is not carried by visible light. (credit: Tatyana Temirbulatova, Flickr)

Diffraction is not only a problem for optical instruments but also for the electromagnetic radiation itself. Any beam of light having a finite diameter D and a wavelength λ exhibits diffraction spreading. The beam spreads out with an angle θ given by the equation $\theta = 1.22\frac{\lambda}{D}$. Take, for example, a laser beam made of rays as parallel as possible (angles between rays as close to $\theta = 0^{\circ}$ as possible) instead spreads out at an angle $\theta = 1.22\ \lambda/D$, where D is the diameter of the beam and λ is its wavelength. This spreading is impossible to observe for a flashlight, because its beam is not very parallel to start with. However, for long-distance transmission of laser beams or microwave signals, diffraction spreading can be significant (see [link]). To avoid this, we can increase D. This is done for laser light sent to the Moon to measure its distance from the Earth. The laser beam is expanded through a telescope to make D much larger and θ smaller.



The beam produced by this microwave transmission antenna will spread out at a minimum angle $\theta = 1.22\ \lambda / D$ due to diffraction. It is impossible to produce a near-parallel beam, because the beam has a limited diameter.

In most biology laboratories, resolution is presented when the use of the microscope is introduced. The ability of a lens to produce sharp images of two closely spaced point objects is called resolution. The smaller the distance x by which two objects can be separated and still be seen as distinct, the greater the resolution. The resolving power of a lens is defined as that distance x. An expression for resolving power is obtained from the

Rayleigh criterion. In [link](a) we have two point objects separated by a distance x. According to the Rayleigh criterion, resolution is possible when the minimum angular separation is

Equation:

$$\theta = 1.22\frac{\lambda}{D} = \frac{x}{d},$$

where d is the distance between the specimen and the objective lens, and we have used the small angle approximation (i.e., we have assumed that x is much smaller than d), so that $\tan\theta \approx \sin\theta \approx \theta$.

Therefore, the resolving power is

Equation:

$$x = 1.22\frac{\lambda d}{D}.$$

Another way to look at this is by re-examining the concept of Numerical Aperture (NA) discussed in Microscopes. There, NA is a measure of the maximum acceptance angle at which the fiber will take light and still contain it within the fiber. [link](b) shows a lens and an object at point P. The NA here is a measure of the ability of the lens to gather light and resolve fine detail. The angle subtended by the lens at its focus is defined to be $\theta = 2\alpha$. From the figure and again using the small angle approximation, we can write

Equation:

$$\sin\alpha = \frac{D/2}{d} = \frac{D}{2d}.$$

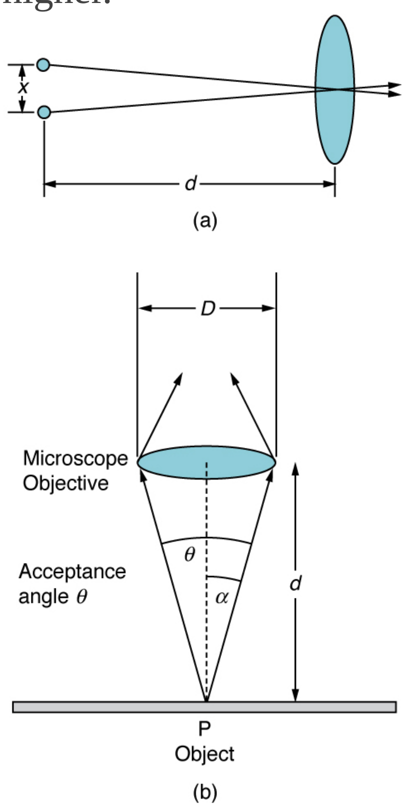
The NA for a lens is $NA = n\sin\alpha$, where n is the index of refraction of the medium between the objective lens and the object at point P.

From this definition for NA, we can see that

Equation:

$$x = 1.22\frac{\lambda d}{D} = 1.22\frac{\lambda}{2\ \sin\ \alpha} = 0.61\frac{\lambda n}{\text{NA}}.$$

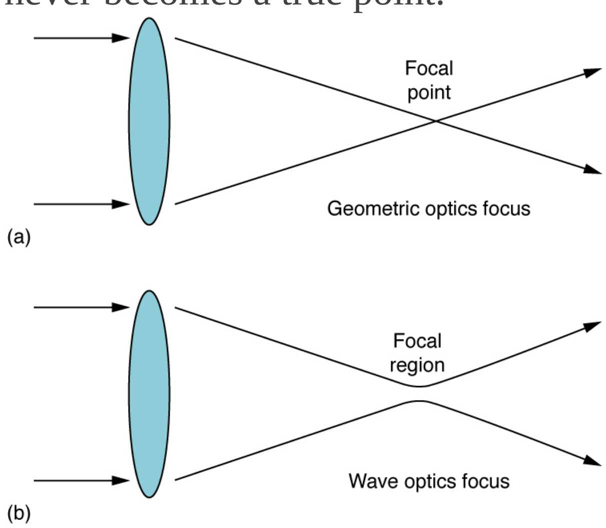
In a microscope, NA is important because it relates to the resolving power of a lens. A lens with a large NA will be able to resolve finer details. Lenses with larger NA will also be able to collect more light and so give a brighter image. Another way to describe this situation is that the larger the NA, the larger the cone of light that can be brought into the lens, and so more of the diffraction modes will be collected. Thus the microscope has more information to form a clear image, and so its resolving power will be higher.



(a) Two points separated by at distance x and a positioned a distance d away from the objective. (credit: Infopro,

Wikimedia Commons) (b) Terms and symbols used in discussion of resolving power for a lens and an object at point P. (credit: Infopro, Wikimedia Commons)

One of the consequences of diffraction is that the focal point of a beam has a finite width and intensity distribution. Consider focusing when only considering geometric optics, shown in [link](a). The focal point is infinitely small with a huge intensity and the capacity to incinerate most samples irrespective of the NA of the objective lens. For wave optics, due to diffraction, the focal point spreads to become a focal spot (see [link](b)) with the size of the spot decreasing with increasing NA. Consequently, the intensity in the focal spot increases with increasing NA. The higher the NA, the greater the chances of photodegrading the specimen. However, the spot never becomes a true point.



(a) In geometric optics, the focus is a point, but it is not

physically possible to produce such a point because it implies infinite intensity. (b) In wave optics, the focus is an extended region.

Section Summary

- Diffraction limits resolution.
- For a circular aperture, lens, or mirror, the Rayleigh criterion states that two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other.
- This occurs for two point objects separated by the angle $\theta = 1.22\frac{\lambda}{D}$, where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc. This equation also gives the angular spreading of a source of light having a diameter D.

Conceptual Questions

Exercise:

Problem:

A beam of light always spreads out. Why can a beam not be created with parallel rays to prevent spreading? Why can lenses, mirrors, or apertures not be used to correct the spreading?

Problems & Exercises

Exercise:

Problem:

The 300-m-diameter Arecibo radio telescope pictured in [link] detects radio waves with a 4.00 cm average wavelength.

(a) What is the angle between two just-resolvable point sources for this telescope?

(b) How close together could these point sources be at the 2 million light year distance of the Andromeda galaxy?

Solution:

(a) 1.63×10^{-4} rad

(b) 326 ly

Exercise:

Problem:

Assuming the angular resolution found for the Hubble Telescope in [link], what is the smallest detail that could be observed on the Moon?

Exercise:

Problem:

Diffraction spreading for a flashlight is insignificant compared with other limitations in its optics, such as spherical aberrations in its mirror. To show this, calculate the minimum angular spreading of a flashlight beam that is originally 5.00 cm in diameter with an average wavelength of 600 nm.

Solution:

1.46×10^{-5} rad

Exercise:

Problem:

(a) What is the minimum angular spread of a 633-nm wavelength He-Ne laser beam that is originally 1.00 mm in diameter?

(b) If this laser is aimed at a mountain cliff 15.0 km away, how big will the illuminated spot be?

(c) How big a spot would be illuminated on the Moon, neglecting atmospheric effects? (This might be done to hit a corner reflector to measure the round-trip time and, hence, distance.) Explicitly show how you follow the steps in Problem-Solving Strategies for Wave Optics.

Exercise:

Problem:

A telescope can be used to enlarge the diameter of a laser beam and limit diffraction spreading. The laser beam is sent through the telescope in opposite the normal direction and can then be projected onto a satellite or the Moon.

(a) If this is done with the Mount Wilson telescope, producing a 2.54-m-diameter beam of 633-nm light, what is the minimum angular spread of the beam?

(b) Neglecting atmospheric effects, what is the size of the spot this beam would make on the Moon, assuming a lunar distance of 3.84×10^8 m?

Solution:

(a) 3.04×10^{-7} rad

(b) Diameter of 235 m

Exercise:

Problem:

The limit to the eye's acuity is actually related to diffraction by the pupil.

(a) What is the angle between two just-resolvable points of light for a 3.00-mm-diameter pupil, assuming an average wavelength of 550 nm?

(b) Take your result to be the practical limit for the eye. What is the greatest possible distance a car can be from you if you can resolve its two headlights, given they are 1.30 m apart?

(c) What is the distance between two just-resolvable points held at an arm's length (0.800 m) from your eye?

(d) How does your answer to (c) compare to details you normally observe in everyday circumstances?

Exercise:

Problem:

What is the minimum diameter mirror on a telescope that would allow you to see details as small as 5.00 km on the Moon some 384,000 km away? Assume an average wavelength of 550 nm for the light received.

Solution:

5.15 cm

Exercise:

Problem:

You are told not to shoot until you see the whites of their eyes. If the eyes are separated by 6.5 cm and the diameter of your pupil is 5.0 mm, at what distance can you resolve the two eyes using light of wavelength 555 nm?

Exercise:

Problem:

(a) The planet Pluto and its Moon Charon are separated by 19,600 km. Neglecting atmospheric effects, should the 5.08-m-diameter Mount Palomar telescope be able to resolve these bodies when they are 4.50×10^9 km from Earth? Assume an average wavelength of 550 nm.

(b) In actuality, it is just barely possible to discern that Pluto and Charon are separate bodies using an Earth-based telescope. What are the reasons for this?

Solution:

(a) Yes. Should easily be able to discern.

(b) The fact that it is just barely possible to discern that these are separate bodies indicates the severity of atmospheric aberrations.

Exercise:

Problem:

The headlights of a car are 1.3 m apart. What is the maximum distance at which the eye can resolve these two headlights? Take the pupil diameter to be 0.40 cm.

Exercise:

Problem:

When dots are placed on a page from a laser printer, they must be close enough so that you do not see the individual dots of ink. To do this, the separation of the dots must be less than Raleigh's criterion. Take the pupil of the eye to be 3.0 mm and the distance from the paper to the eye of 35 cm; find the minimum separation of two dots such that they cannot be resolved. How many dots per inch (dpi) does this correspond to?

Exercise:

Problem: Unreasonable Results

An amateur astronomer wants to build a telescope with a diffraction limit that will allow him to see if there are people on the moons of Jupiter.

(a) What diameter mirror is needed to be able to see 1.00 m detail on a Jovian Moon at a distance of 7.50×10^8 km from Earth? The wavelength of light averages 600 nm.

(b) What is unreasonable about this result?

(c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: Construct Your Own Problem

Consider diffraction limits for an electromagnetic wave interacting with a circular object. Construct a problem in which you calculate the limit of angular resolution with a device, using this circular object (such as a lens, mirror, or antenna) to make observations. Also calculate the limit to spatial resolution (such as the size of features observable on the Moon) for observations at a specific distance from the device. Among the things to be considered are the wavelength of electromagnetic radiation used, the size of the circular object, and the distance to the system or phenomenon being observed.

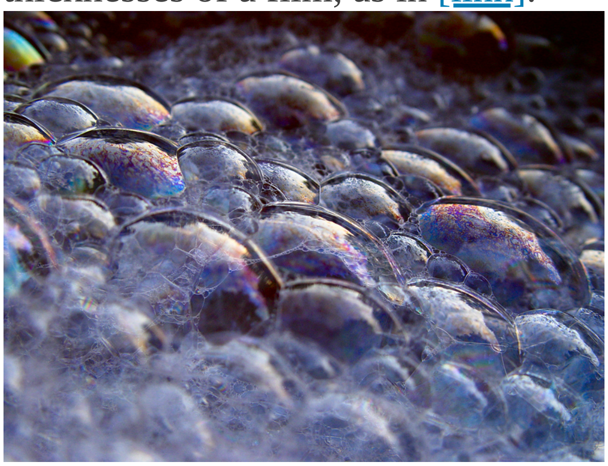
Glossary

Rayleigh criterion
: two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other

Thin Film Interference

- Discuss the rainbow formation by thin films.

The bright colors seen in an oil slick floating on water or in a sunlit soap bubble are caused by interference. The brightest colors are those that interfere constructively. This interference is between light reflected from different surfaces of a thin film; thus, the effect is known as **thin film interference**. As noticed before, interference effects are most prominent when light interacts with something having a size similar to its wavelength. A thin film is one having a thickness t smaller than a few times the wavelength of light, λ. Since color is associated indirectly with λ and since all interference depends in some way on the ratio of λ to the size of the object involved, we should expect to see different colors for different thicknesses of a film, as in [link].

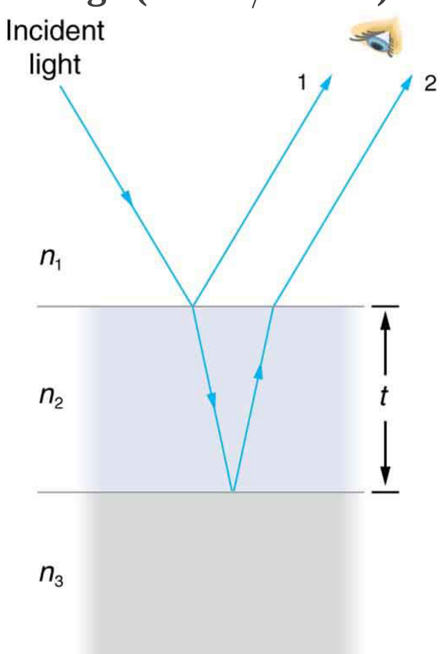


These soap bubbles exhibit brilliant colors when exposed to sunlight. (credit: Scott Robinson, Flickr)

What causes thin film interference? [link] shows how light reflected from the top and bottom surfaces of a film can interfere. Incident light is only partially reflected from the top surface of the film (ray 1). The remainder enters the film and is itself partially reflected from the bottom surface. Part

of the light reflected from the bottom surface can emerge from the top of the film (ray 2) and interfere with light reflected from the top (ray 1). Since the ray that enters the film travels a greater distance, it may be in or out of phase with the ray reflected from the top. However, consider for a moment, again, the bubbles in [link]. The bubbles are darkest where they are thinnest. Furthermore, if you observe a soap bubble carefully, you will note it gets dark at the point where it breaks. For very thin films, the difference in path lengths of ray 1 and ray 2 in [link] is negligible; so why should they interfere destructively and not constructively? The answer is that a phase change can occur upon reflection. The rule is as follows:

When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180° phase change (or a $\lambda/2$ shift) occurs.



Light striking a thin film is partially reflected (ray 1) and partially refracted at the top surface. The refracted ray is partially reflected at the bottom surface

and emerges as ray 2. These rays will interfere in a way that depends on the thickness of the film and the indices of refraction of the various media.

If the film in [link] is a soap bubble (essentially water with air on both sides), then there is a $\lambda/2$ shift for ray 1 and none for ray 2. Thus, when the film is very thin, the path length difference between the two rays is negligible, they are exactly out of phase, and destructive interference will occur at all wavelengths and so the soap bubble will be dark here.

The thickness of the film relative to the wavelength of light is the other crucial factor in thin film interference. Ray 2 in [link] travels a greater distance than ray 1. For light incident perpendicular to the surface, ray 2 travels a distance approximately $2t$ farther than ray 1. When this distance is an integral or half-integral multiple of the wavelength in the medium ($\lambda_n = \lambda/n$, where λ is the wavelength in vacuum and n is the index of refraction), constructive or destructive interference occurs, depending also on whether there is a phase change in either ray.

Example:
Calculating Non-reflective Lens Coating Using Thin Film Interference
Sophisticated cameras use a series of several lenses. Light can reflect from the surfaces of these various lenses and degrade image clarity. To limit these reflections, lenses are coated with a thin layer of magnesium fluoride that causes destructive thin film interference. What is the thinnest this film can be, if its index of refraction is 1.38 and it is designed to limit the reflection of 550-nm light, normally the most intense visible wavelength? The index of refraction of glass is 1.52.

Strategy

Refer to [link] and use $n_1 = 100$ for air, $n_2 = 1.38$, and $n_3 = 1.52$. Both ray 1 and ray 2 will have a $\lambda/2$ shift upon reflection. Thus, to obtain destructive interference, ray 2 will need to travel a half wavelength farther than ray 1. For rays incident perpendicularly, the path length difference is $2t$.

Solution

To obtain destructive interference here,

Equation:

$$2t = \frac{\lambda_{n_2}}{2},$$

where λ_{n_2} is the wavelength in the film and is given by $\lambda_{n_2} = \frac{\lambda}{n_2}$.

Thus,

Equation:

$$2t = \frac{\lambda/n_2}{2}.$$

Solving for t and entering known values yields

Equation:

$$\begin{aligned} t &= \frac{\lambda/n_2}{4} = \frac{(550\text{ nm})/1.38}{4} \\ &= 99.6\text{ nm}. \end{aligned}$$

Discussion

Films such as the one in this example are most effective in producing destructive interference when the thinnest layer is used, since light over a broader range of incident angles will be reduced in intensity. These films are called non-reflective coatings; this is only an approximately correct description, though, since other wavelengths will only be partially cancelled. Non-reflective coatings are used in car windows and sunglasses.

Thin film interference is most constructive or most destructive when the path length difference for the two rays is an integral or half-integral wavelength, respectively. That is, for rays incident perpendicularly, $2t = \lambda_n \;\; 2\lambda_n \;\; 3\lambda_n \;\; \ldots$ or $2t = \lambda_n/2, 3\lambda_n/2, 5\lambda_n/2, \ldots$. To know whether interference is constructive or destructive, you must also determine if there is a phase change upon reflection. Thin film interference thus depends on film thickness, the wavelength of light, and the refractive indices. For white light incident on a film that varies in thickness, you will observe rainbow colors of constructive interference for various wavelengths as the thickness varies.

Example:
Soap Bubbles: More Than One Thickness can be Constructive
(a) What are the three smallest thicknesses of a soap bubble that produce constructive interference for red light with a wavelength of 650 nm? The index of refraction of soap is taken to be the same as that of water. (b) What three smallest thicknesses will give destructive interference?

Strategy and Concept

Use [link] to visualize the bubble. Note that $n_1 = n_3 = 1.00$ for air, and $n_2 = 1.333$ for soap (equivalent to water). There is a $\lambda/2$ shift for ray 1 reflected from the top surface of the bubble, and no shift for ray 2 reflected from the bottom surface. To get constructive interference, then, the path length difference $(2t)$ must be a half-integral multiple of the wavelength—the first three being $\lambda_n/2$, $3\lambda_n/2$, and $5\lambda_n/2$. To get destructive interference, the path length difference must be an integral multiple of the wavelength—the first three being 0, λ_n, and $2\lambda_n$.

Solution for (a)

Constructive interference occurs here when

Equation:

$$2t_{\text{c}} = \frac{\lambda_n}{2} \;\; \frac{3\lambda_n}{2} \;\; \frac{5\lambda_n}{2}, \ldots .$$

The smallest constructive thickness t_{c} thus is

Equation:

$$\begin{aligned} t_{\mathrm{c}} &= \frac{\lambda_n}{4} = \frac{\lambda/n}{4} = \frac{(650\text{ nm})/1.333}{4} \\ &= 122\text{ nm}. \end{aligned}$$

The next thickness that gives constructive interference is $t\prime_{\mathrm{c}} = 3\lambda_n/4$, so that

Equation:

$$t\prime_{\mathrm{c}} = 366\text{ nm}.$$

Finally, the third thickness producing constructive interference is $t\prime\prime_{\mathrm{c}} \leq 5\lambda_n/4$, so that

Equation:

$$t\prime\prime_{\mathrm{c}} = 610\text{ nm}.$$

Solution for (b)

For *destructive interference*, the path length difference here is an integral multiple of the wavelength. The first occurs for zero thickness, since there is a phase change at the top surface. That is,

Equation:

$$t_{\mathrm{d}} = 0.$$

The first non-zero thickness producing destructive interference is

Equation:

$$2t\prime_{\mathrm{d}} = \lambda_n.$$

Substituting known values gives

Equation:

$$\begin{aligned} t\prime_{\mathrm{d}} &= \frac{\lambda_n}{2} = \frac{\lambda/n}{2} = \frac{(650\text{ nm})/1.333}{2} \\ &= 244\text{ nm}. \end{aligned}$$

Finally, the third destructive thickness is $2t\prime\prime_{\mathrm{d}} = 2\lambda_n$, so that

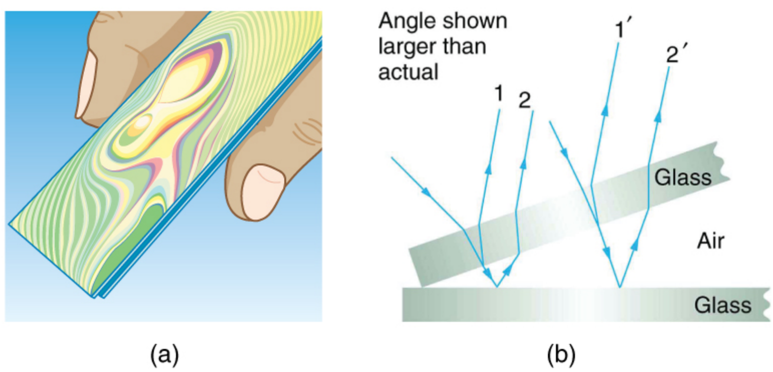
Equation:

$$\begin{aligned} t''_{\mathrm{d}} &= \lambda_n = \frac{\lambda}{n} = \frac{650 \text{ nm}}{1.333} \\ &= 488 \text{ nm}. \end{aligned}$$

Discussion

If the bubble was illuminated with pure red light, we would see bright and dark bands at very uniform increases in thickness. First would be a dark band at 0 thickness, then bright at 122 nm thickness, then dark at 244 nm, bright at 366 nm, dark at 488 nm, and bright at 610 nm. If the bubble varied smoothly in thickness, like a smooth wedge, then the bands would be evenly spaced.

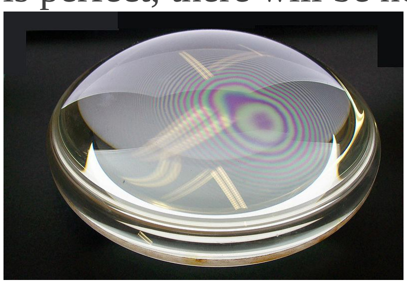
Another example of thin film interference can be seen when microscope slides are separated (see [link]). The slides are very flat, so that the wedge of air between them increases in thickness very uniformly. A phase change occurs at the second surface but not the first, and so there is a dark band where the slides touch. The rainbow colors of constructive interference repeat, going from violet to red again and again as the distance between the slides increases. As the layer of air increases, the bands become more difficult to see, because slight changes in incident angle have greater effects on path length differences. If pure-wavelength light instead of white light is used, then bright and dark bands are obtained rather than repeating rainbow colors.



(a) The rainbow color bands are produced by thin film interference in the air between the two glass slides. (b) Schematic of the

paths taken by rays in the wedge of air between the slides.

An important application of thin film interference is found in the manufacturing of optical instruments. A lens or mirror can be compared with a master as it is being ground, allowing it to be shaped to an accuracy of less than a wavelength over its entire surface. [link] illustrates the phenomenon called Newton’s rings, which occurs when the plane surfaces of two lenses are placed together. (The circular bands are called Newton’s rings because Isaac Newton described them and their use in detail. Newton did not discover them; Robert Hooke did, and Newton did not believe they were due to the wave character of light.) Each successive ring of a given color indicates an increase of only one wavelength in the distance between the lens and the blank, so that great precision can be obtained. Once the lens is perfect, there will be no rings.



“Newton's rings” interference fringes are produced when two plano-convex lenses are placed together with their plane surfaces in contact. The rings are created by interference between the light reflected off the two surfaces as a result of a slight

gap between them, indicating that these surfaces are not precisely plane but are slightly convex. (credit: Ulf Seifert, Wikimedia Commons)

The wings of certain moths and butterflies have nearly iridescent colors due to thin film interference. In addition to pigmentation, the wing's color is affected greatly by constructive interference of certain wavelengths reflected from its film-coated surface. Car manufacturers are offering special paint jobs that use thin film interference to produce colors that change with angle. This expensive option is based on variation of thin film path length differences with angle. Security features on credit cards, banknotes, driving licenses and similar items prone to forgery use thin film interference, diffraction gratings, or holograms. Australia led the way with dollar bills printed on polymer with a diffraction grating security feature making the currency difficult to forge. Other countries such as New Zealand and Taiwan are using similar technologies, while the United States currency includes a thin film interference effect.

Note:
Making Connections: Take-Home Experiment—Thin Film Interference
One feature of thin film interference and diffraction gratings is that the pattern shifts as you change the angle at which you look or move your head. Find examples of thin film interference and gratings around you. Explain how the patterns change for each specific example. Find examples where the thickness changes giving rise to changing colors. If you can find two microscope slides, then try observing the effect shown in [link]. Try separating one end of the two slides with a hair or maybe a thin piece of paper and observe the effect.

Problem-Solving Strategies for Wave Optics

Step 1. *Examine the situation to determine that interference is involved*. Identify whether slits or thin film interference are considered in the problem.

Step 2. *If slits are involved*, note that diffraction gratings and double slits produce very similar interference patterns, but that gratings have narrower (sharper) maxima. Single slit patterns are characterized by a large central maximum and smaller maxima to the sides.

Step 3. *If thin film interference is involved, take note of the path length difference between the two rays that interfere*. Be certain to use the wavelength in the medium involved, since it differs from the wavelength in vacuum. Note also that there is an additional $\lambda/2$ phase shift when light reflects from a medium with a greater index of refraction.

Step 4. *Identify exactly what needs to be determined in the problem (identify the unknowns)*. A written list is useful. Draw a diagram of the situation. Labeling the diagram is useful.

Step 5. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns)*.

Step 6. *Solve the appropriate equation for the quantity to be determined (the unknown), and enter the knowns*. Slits, gratings, and the Rayleigh limit involve equations.

Step 7. *For thin film interference, you will have constructive interference for a total shift that is an integral number of wavelengths. You will have destructive interference for a total shift of a half-integral number of wavelengths*. Always keep in mind that crest to crest is constructive whereas crest to trough is destructive.

Step 8. *Check to see if the answer is reasonable: Does it make sense?* Angles in interference patterns cannot be greater than 90º, for example.

Section Summary

- Thin film interference occurs between the light reflected from the top and bottom surfaces of a film. In addition to the path length difference, there can be a phase change.
- When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180° phase change (or a $\lambda/2$ shift) occurs.

Conceptual Questions

Exercise:

Problem:

What effect does increasing the wedge angle have on the spacing of interference fringes? If the wedge angle is too large, fringes are not observed. Why?

Exercise:

Problem:

How is the difference in paths taken by two originally in-phase light waves related to whether they interfere constructively or destructively? How can this be affected by reflection? By refraction?

Exercise:

Problem:

Is there a phase change in the light reflected from either surface of a contact lens floating on a person’s tear layer? The index of refraction of the lens is about 1.5, and its top surface is dry.

Exercise:

Problem:

In placing a sample on a microscope slide, a glass cover is placed over a water drop on the glass slide. Light incident from above can reflect from the top and bottom of the glass cover and from the glass slide below the water drop. At which surfaces will there be a phase change in the reflected light?

Exercise:

Problem:

Answer the above question if the fluid between the two pieces of crown glass is carbon disulfide.

Exercise:

Problem:

While contemplating the food value of a slice of ham, you notice a rainbow of color reflected from its moist surface. Explain its origin.

Exercise:

Problem:

An inventor notices that a soap bubble is dark at its thinnest and realizes that destructive interference is taking place for all wavelengths. How could she use this knowledge to make a non-reflective coating for lenses that is effective at all wavelengths? That is, what limits would there be on the index of refraction and thickness of the coating? How might this be impractical?

Exercise:

Problem:

A non-reflective coating like the one described in [link] works ideally for a single wavelength and for perpendicular incidence. What happens for other wavelengths and other incident directions? Be specific.

Exercise:

Problem:

Why is it much more difficult to see interference fringes for light reflected from a thick piece of glass than from a thin film? Would it be easier if monochromatic light were used?

Problems & Exercises

Exercise:

Problem:

A soap bubble is 100 nm thick and illuminated by white light incident perpendicular to its surface. What wavelength and color of visible light is most constructively reflected, assuming the same index of refraction as water?

Solution:

532 nm (green)

Exercise:

Problem:

An oil slick on water is 120 nm thick and illuminated by white light incident perpendicular to its surface. What color does the oil appear (what is the most constructively reflected wavelength), given its index of refraction is 1.40?

Exercise:

Problem:

Calculate the minimum thickness of an oil slick on water that appears blue when illuminated by white light perpendicular to its surface. Take the blue wavelength to be 470 nm and the index of refraction of oil to be 1.40.

Solution:

83.9 nm

Exercise:

Problem:

Find the minimum thickness of a soap bubble that appears red when illuminated by white light perpendicular to its surface. Take the wavelength to be 680 nm, and assume the same index of refraction as water.

Exercise:

Problem:

A film of soapy water ($n = 1.33$) on top of a plastic cutting board has a thickness of 233 nm. What color is most strongly reflected if it is illuminated perpendicular to its surface?

Solution:

620 nm (orange)

Exercise:

Problem:

What are the three smallest non-zero thicknesses of soapy water ($n = 1.33$) on Plexiglas if it appears green (constructively reflecting 520-nm light) when illuminated perpendicularly by white light? Explicitly show how you follow the steps in Problem Solving Strategies for Wave Optics.

Exercise:

Problem:

Suppose you have a lens system that is to be used primarily for 700-nm red light. What is the second thinnest coating of fluorite (magnesium fluoride) that would be non-reflective for this wavelength?

Solution:

380 nm

Exercise:

Problem:

(a) As a soap bubble thins it becomes dark, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the bubble can be and appear dark at all visible wavelengths? Assume the same index of refraction as water. (b) Discuss the fragility of the film considering the thickness found.

Exercise:

Problem:

A film of oil on water will appear dark when it is very thin, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the oil can be and appear dark at all visible wavelengths? Oil has an index of refraction of 1.40.

Solution:

33.9 nm

Exercise:

Problem:

[link] shows two glass slides illuminated by pure-wavelength light incident perpendicularly. The top slide touches the bottom slide at one end and rests on a 0.100-mm-diameter hair at the other end, forming a wedge of air. (a) How far apart are the dark bands, if the slides are 7.50 cm long and 589-nm light is used? (b) Is there any difference if the slides are made from crown or flint glass? Explain.

Exercise:

Problem:

[link] shows two 7.50-cm-long glass slides illuminated by pure 589-nm wavelength light incident perpendicularly. The top slide touches the bottom slide at one end and rests on some debris at the other end, forming a wedge of air. How thick is the debris, if the dark bands are 1.00 mm apart?

Solution:

4.42×10^{-5} m

Exercise:

Problem: Repeat [link], but take the light to be incident at a 45^{o} angle.

Exercise:

Problem: Repeat [link], but take the light to be incident at a 45^{o} angle.

Solution:

The oil film will appear black, since the reflected light is not in the visible part of the spectrum.

Exercise:

Problem: **Unreasonable Results**

To save money on making military aircraft invisible to radar, an inventor decides to coat them with a non-reflective material having an index of refraction of 1.20, which is between that of air and the surface of the plane. This, he reasons, should be much cheaper than designing Stealth bombers. (a) What thickness should the coating be to inhibit the reflection of 4.00-cm wavelength radar? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

thin film interference
: interference between light reflected from different surfaces of a thin film

Polarization

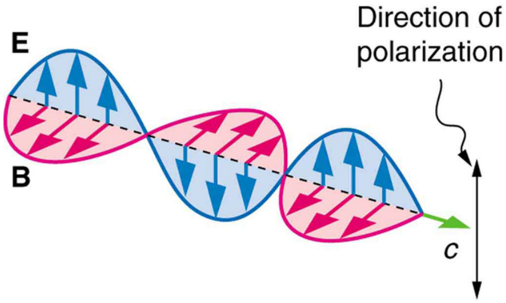
- Discuss the meaning of polarization.
- Discuss the property of optical activity of certain materials.

Polaroid sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass (see [link]). Polaroids have this ability because of a wave characteristic of light called polarization. What is polarization? How is it produced? What are some of its uses? The answers to these questions are related to the wave character of light.

These two photographs of a river show the effect of a polarizing filter in reducing glare in light reflected from the surface of water. Part (b) of this figure was taken with a polarizing filter and part (a) was not. As a result, the reflection of clouds and sky observed in part (a) is not observed in part (b). Polarizing sunglasses are particularly useful on snow and water. (credit: Amithshs, Wikimedia Commons)

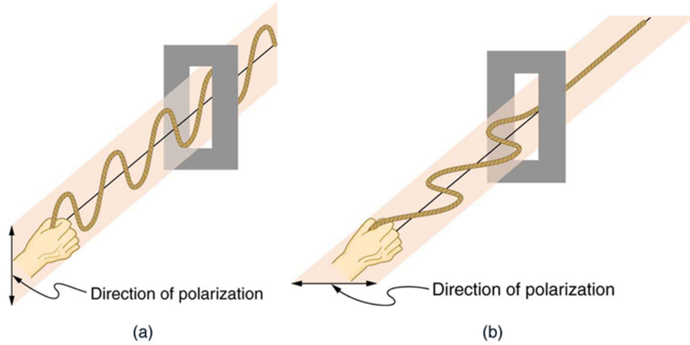
Light is one type of electromagnetic (EM) wave. As noted earlier, EM waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation (see [link]). There are specific directions for the oscillations of the electric and magnetic fields. **Polarization** is the attribute that a wave's oscillations have

a definite direction relative to the direction of propagation of the wave. (This is not the same type of polarization as that discussed for the separation of charges.) Waves having such a direction are said to be **polarized**. For an EM wave, we define the **direction of polarization** to be the direction parallel to the electric field. Thus we can think of the electric field arrows as showing the direction of polarization, as in [link].



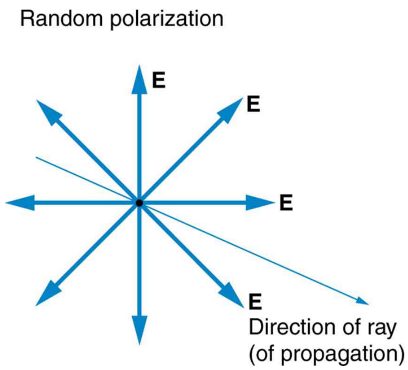
An EM wave, such as light, is a transverse wave. The electric and magnetic fields are perpendicular to the direction of propagation.

To examine this further, consider the transverse waves in the ropes shown in [link]. The oscillations in one rope are in a vertical plane and are said to be **vertically polarized**. Those in the other rope are in a horizontal plane and are **horizontally polarized**. If a vertical slit is placed on the first rope, the waves pass through. However, a vertical slit blocks the horizontally polarized waves. For EM waves, the direction of the electric field is analogous to the disturbances on the ropes.

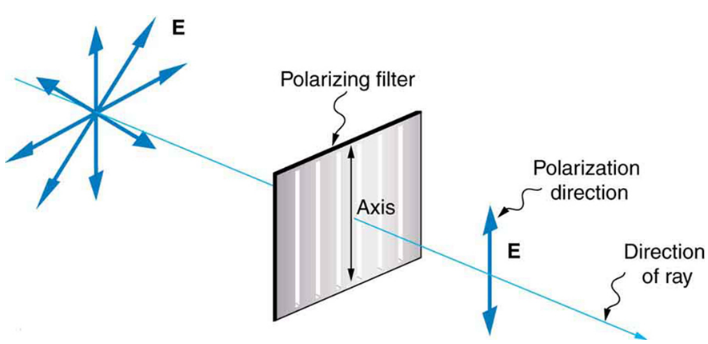


The transverse oscillations in one rope are in a vertical plane, and those in the other rope are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

The Sun and many other light sources produce waves that are randomly polarized (see [link]). Such light is said to be **unpolarized** because it is composed of many waves with all possible directions of polarization. Polaroid materials, invented by the founder of Polaroid Corporation, Edwin Land, act as a *polarizing* slit for light, allowing only polarization in one direction to pass through. Polarizing filters are composed of long molecules aligned in one direction. Thinking of the molecules as many slits, analogous to those for the oscillating ropes, we can understand why only light with a specific polarization can get through. The **axis of a polarizing filter** is the direction along which the filter passes the electric field of an EM wave (see [link]).



The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. Since the light is unpolarized, the arrows point in all directions.



A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM

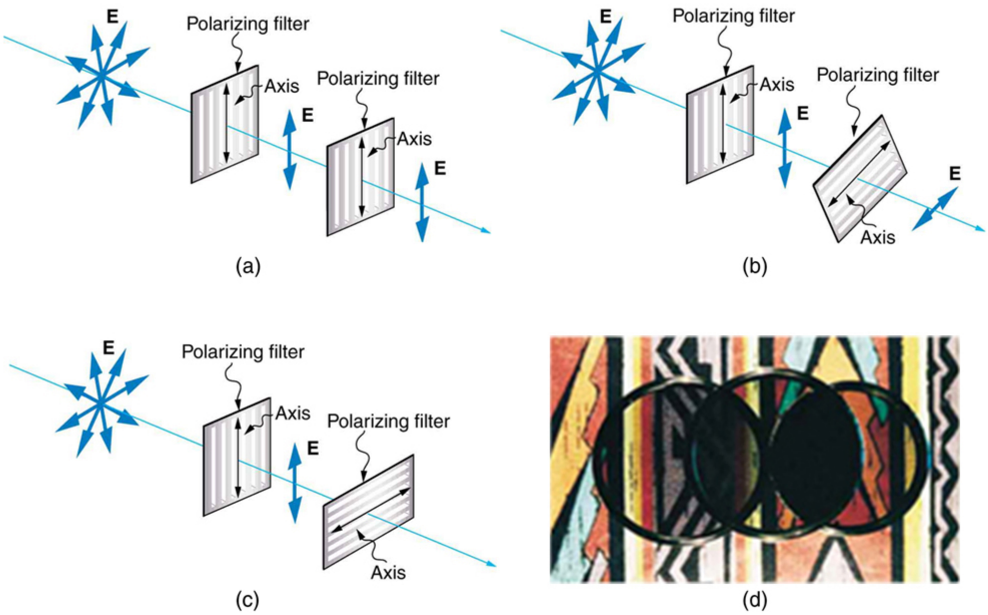
wave is defined to be the direction of
its electric field.

[link] shows the effect of two polarizing filters on originally unpolarized light. The first filter polarizes the light along its axis. When the axes of the first and second filters are aligned (parallel), then all of the polarized light passed by the first filter is also passed by the second. If the second polarizing filter is rotated, only the component of the light parallel to the second filter's axis is passed. When the axes are perpendicular, no light is passed by the second.

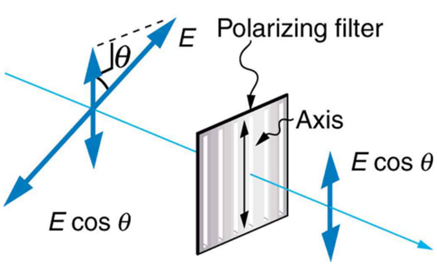
Only the component of the EM wave parallel to the axis of a filter is passed. Let us call the angle between the direction of polarization and the axis of a filter . If the electric field has an amplitude , then the transmitted part of the wave has an amplitude (see [link]). Since the intensity of a wave is proportional to its amplitude squared, the intensity of the transmitted wave is related to the incident wave by

Equation:

where is the intensity of the polarized wave before passing through the filter. (The above equation is known as Malus's law.)



The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second is rotated, only part of the light is passed. (c) When the second is perpendicular to the first, no light is passed. (d) In this photograph, a polarizing filter is placed above two others. Its axis is perpendicular to the filter on the right (dark area) and parallel to the filter on the left (lighter area). (credit: P.P. Urone)



A polarizing filter transmits only the component of the wave parallel to its

axis, ,
reducing the intensity
of any light not
polarized parallel to
its axis.

Example:

Calculating Intensity Reduction by a Polarizing Filter

What angle is needed between the direction of polarized light and the axis of a polarizing filter to reduce its intensity by ?

Strategy

When the intensity is reduced by , it is or 0.100 times its original value. That is, . Using this information, the equation can be used to solve for the needed angle.

Solution

Solving the equation for and substituting with the relationship between and gives

Equation:

Solving for yields

Equation:

o

Discussion

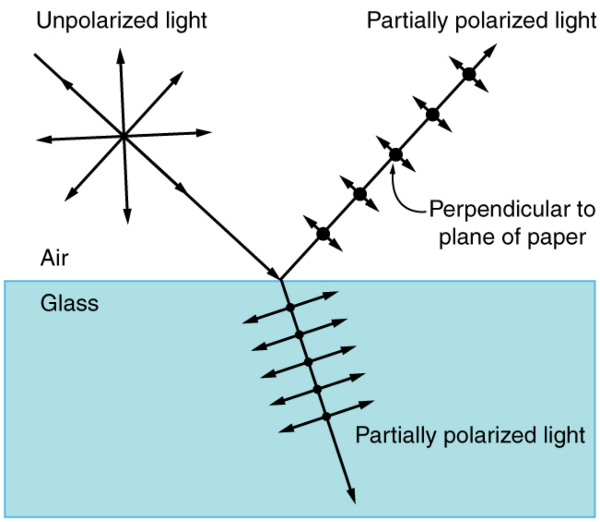
A fairly large angle between the direction of polarization and the filter axis is needed to reduce the intensity to of its original value. This seems reasonable based on experimenting with polarizing films. It is interesting that, at an angle of º, the intensity is reduced to of its original value (as you will show in this section's Problems & Exercises). Note that º

is º from reducing the intensity to zero, and that at an angle of º the intensity is reduced to of its original value (as you will also show in Problems & Exercises), giving evidence of symmetry.

Polarization by Reflection

By now you can probably guess that Polaroid sunglasses cut the glare in reflected light because that light is polarized. You can check this for yourself by holding Polaroid sunglasses in front of you and rotating them while looking at light reflected from water or glass. As you rotate the sunglasses, you will notice the light gets bright and dim, but not completely black. This implies the reflected light is partially polarized and cannot be completely blocked by a polarizing filter.

[link] illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so that *the reflected light is left more horizontally polarized*. The reasons for this phenomenon are beyond the scope of this text, but a convenient mnemonic for remembering this is to imagine the polarization direction to be like an arrow. Vertical polarization would be like an arrow perpendicular to the surface and would be more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and would be more likely to be reflected. Sunglasses with vertical axes would then block more reflected light than unpolarized light from other sources.



Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides bouncing off, whereas arrows striking on their tips go into the surface.

Since the part of the light that is not reflected is refracted, the amount of polarization depends on the indices of refraction of the media involved. It can be shown that **reflected light is completely polarized** at a angle of reflection , given by

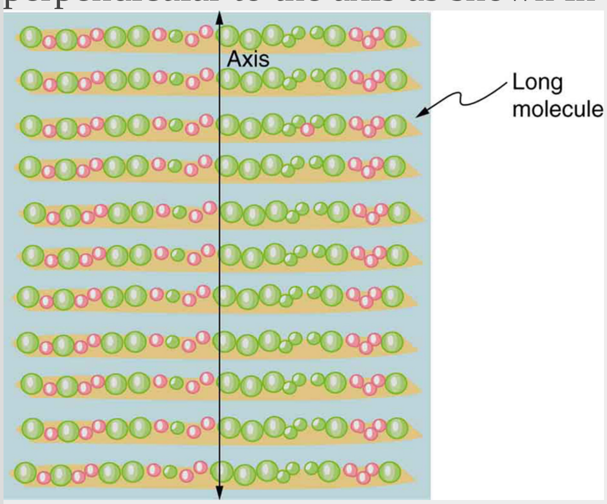
Equation:

where is the medium in which the incident and reflected light travel and is the index of refraction of the medium that forms the interface that reflects the light. This equation is known as **Brewster’s law**, and is known as **Brewster’s angle**, named after the 19th-century Scottish physicist who discovered them.

Note:

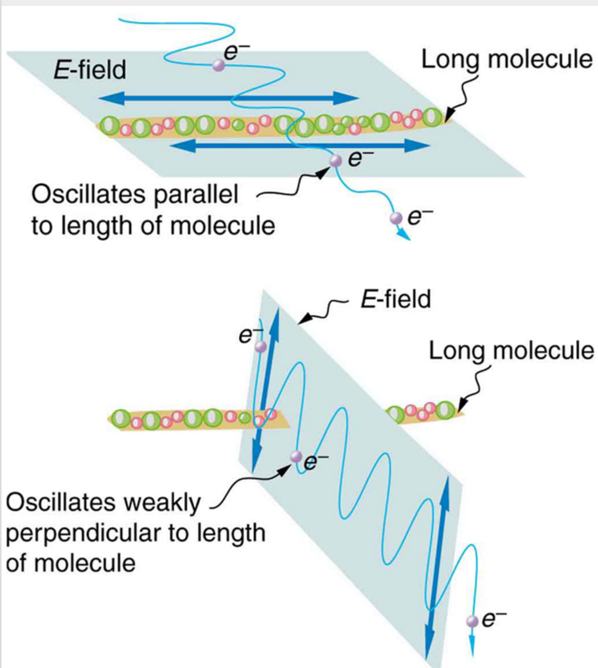
Things Great and Small: Atomic Explanation of Polarizing Filters

Polarizing filters have a polarization axis that acts as a slit. This slit passes electromagnetic waves (often visible light) that have an electric field parallel to the axis. This is accomplished with long molecules aligned perpendicular to the axis as shown in [link].



Long molecules are aligned perpendicular to the axis of a polarizing filter. The component of the electric field in an EM wave perpendicular to these molecules passes through the filter, while the component parallel to the molecules is absorbed.

[link] illustrates how the component of the electric field parallel to the long molecules is absorbed. An electromagnetic wave is composed of oscillating electric and magnetic fields. The electric field is strong compared with the magnetic field and is more effective in exerting force on charges in the molecules. The most affected charged particles are the electrons in the molecules, since electron masses are small. If the electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the fields in the wave and, hence, reduces its intensity. In long molecules, electrons can more easily oscillate parallel to the molecule than in the perpendicular direction. The electrons are bound to the molecule and are more restricted in their movement perpendicular to the molecule. Thus, the electrons can absorb EM waves that have a component of their electric field parallel to the molecule. The electrons are much less responsive to electric fields perpendicular to the molecule and will allow those fields to pass. Thus the axis of the polarizing filter is perpendicular to the length of the molecule.



Artist’s conception of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and

reduces the intensity of the component of the EM wave that is parallel to the molecule.

Example:

Calculating Polarization by Reflection

(a) At what angle will light traveling in air be completely polarized horizontally when reflected from water? (b) From glass?

Strategy

All we need to solve these problems are the indices of refraction. Air has water has and crown glass has . The equation — can be directly applied to find in each case.

Solution for (a)

Putting the known quantities into the equation

Equation:

gives

Equation:

Solving for the angle yields

Equation:

o

Solution for (b)

Similarly, for crown glass and air,

Equation:

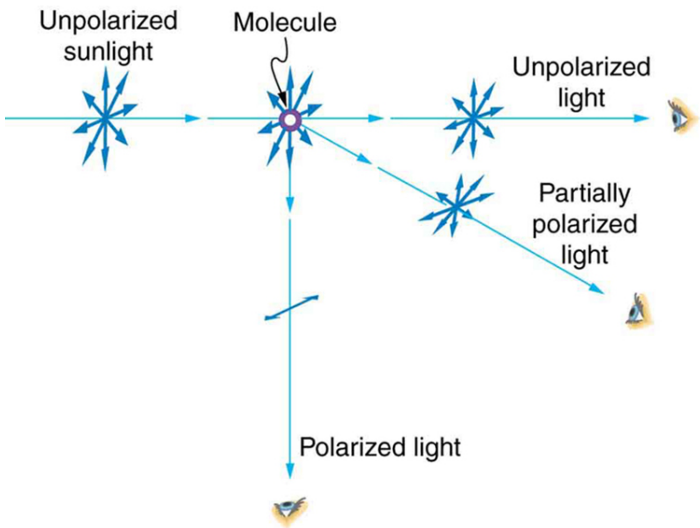
Thus,
Equation:

o

Discussion
Light reflected at these angles could be completely blocked by a good polarizing filter held with its *axis vertical*. Brewster’s angle for water and air are similar to those for glass and air, so that sunglasses are equally effective for light reflected from either water or glass under similar circumstances. Light not reflected is refracted into these media. So at an incident angle equal to Brewster’s angle, the refracted light will be slightly polarized vertically. It will not be completely polarized vertically, because only a small fraction of the incident light is reflected, and so a significant amount of horizontally polarized light is refracted.

Polarization by Scattering

If you hold your Polaroid sunglasses in front of you and rotate them while looking at blue sky, you will see the sky get bright and dim. This is a clear indication that light scattered by air is partially polarized. [link] helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction it is traveling. The electrons then radiate like small antennae. Since they are oscillating perpendicular to the direction of the light ray, they produce EM radiation that is polarized perpendicular to the direction of the ray. When viewing the light along a line perpendicular to the original ray, as in [link], there can be no polarization in the scattered light parallel to the original ray, because that would require the original ray to be a longitudinal wave. Along other directions, a component of the other polarization can be projected along the line of sight, and the scattered light will only be partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.



Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

Photographs of the sky can be darkened by polarizing filters, a trick used by many photographers to make clouds brighter by contrast. Scattering from other particles, such as smoke or dust, can also polarize light. Detecting polarization in scattered EM waves can be a useful analytical tool in determining the scattering source.

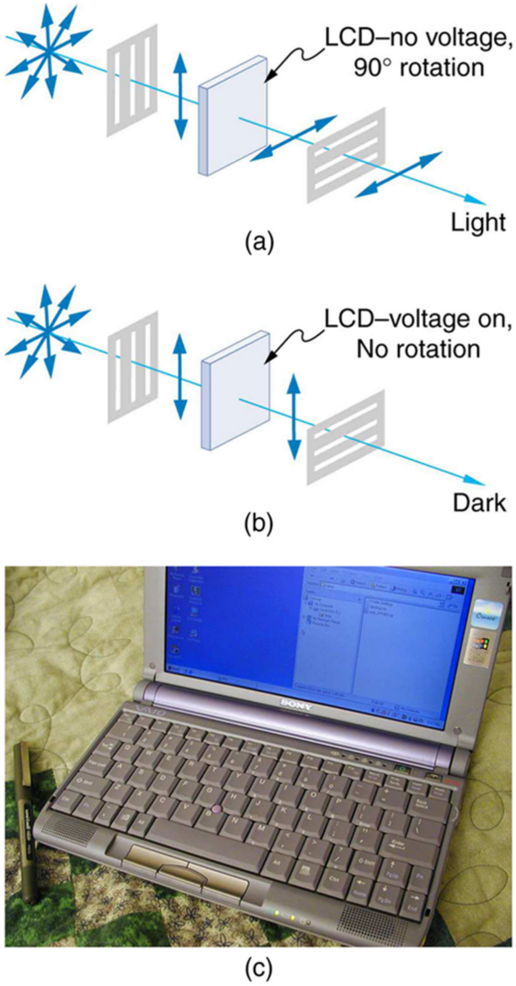
There is a range of optical effects used in sunglasses. Besides being Polaroid, other sunglasses have colored pigments embedded in them, while others use non-reflective or even reflective coatings. A recent development is photochromic lenses, which darken in the sunlight and become clear indoors. Photochromic lenses are embedded with organic microcrystalline molecules that change their properties when exposed to UV in sunlight, but become clear in artificial lighting with no UV.

Note:
Take-Home Experiment: Polarization
Find Polaroid sunglasses and rotate one while holding the other still and look at different surfaces and objects. Explain your observations. What is the difference in angle from when you see a maximum intensity to when you see a minimum intensity? Find a reflective glass surface and do the same. At what angle does the glass need to be oriented to give minimum glare?

Liquid Crystals and Other Polarization Effects in Materials

While you are undoubtedly aware of liquid crystal displays (LCDs) found in watches, calculators, computer screens, cellphones, flat screen televisions, and other myriad places, you may not be aware that they are based on polarization. Liquid crystals are so named because their molecules can be aligned even though they are in a liquid. Liquid crystals have the property that they can rotate the polarization of light passing through them by ^{o}. Furthermore, this property can be turned off by the application of a voltage, as illustrated in [link]. It is possible to manipulate this characteristic quickly and in small well-defined regions to create the contrast patterns we see in so many LCD devices.

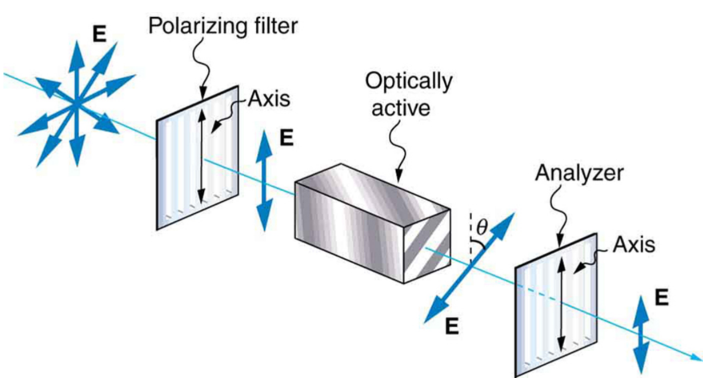
In flat screen LCD televisions, there is a large light at the back of the TV. The light travels to the front screen through millions of tiny units called pixels (picture elements). One of these is shown in [link] (a) and (b). Each unit has three cells, with red, blue, or green filters, each controlled independently. When the voltage across a liquid crystal is switched off, the liquid crystal passes the light through the particular filter. One can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.



(a) Polarized light is rotated º by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the original polarization direction. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making the region dark in comparison with its surroundings. (c) LCDs

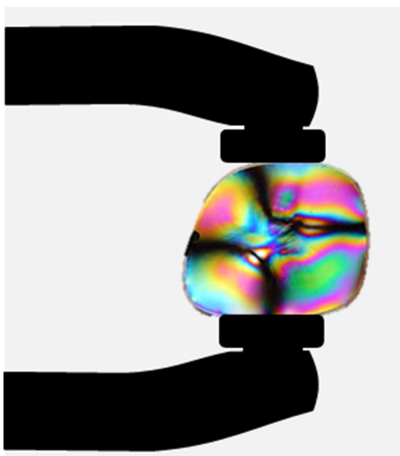
can be made color specific, small, and fast enough to use in laptop computers and TVs. (credit: Jon Sullivan)

Many crystals and solutions rotate the plane of polarization of light passing through them. Such substances are said to be **optically active**. Examples include sugar water, insulin, and collagen (see [link]). In addition to depending on the type of substance, the amount and direction of rotation depends on a number of factors. Among these is the concentration of the substance, the distance the light travels through it, and the wavelength of light. Optical activity is due to the asymmetric shape of molecules in the substance, such as being helical. Measurements of the rotation of polarized light passing through substances can thus be used to measure concentrations, a standard technique for sugars. It can also give information on the shapes of molecules, such as proteins, and factors that affect their shapes, such as temperature and pH.



Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

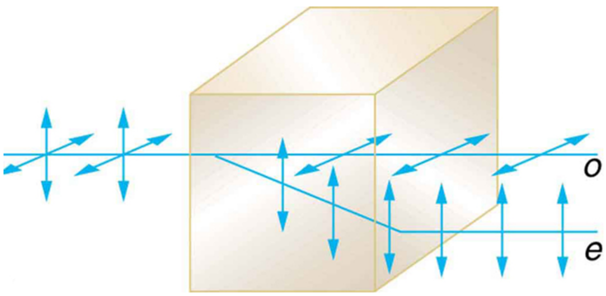
Glass and plastic become optically active when stressed; the greater the stress, the greater the effect. Optical stress analysis on complicated shapes can be performed by making plastic models of them and observing them through crossed filters, as seen in [link]. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.



Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: Infopro, Wikimedia Commons)

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two. Such crystals are said to be **birefringent** (see [link]). Each of the separated rays has a specific polarization. One behaves normally and is called the ordinary ray, whereas the other does not obey Snell’s law and is called the

extraordinary ray. Birefringent crystals can be used to produce polarized beams from unpolarized light. Some birefringent materials preferentially absorb one of the polarizations. These materials are called dichroic and can produce polarization by this preferential absorption. This is fundamentally how polarizing filters and other polarizers work. The interested reader is invited to further pursue the numerous properties of materials related to polarization.



Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two. The ordinary ray behaves as expected, but the extraordinary ray does not obey Snell's law.

Section Summary

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave.
- EM waves are transverse waves that may be polarized.
- The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.
- Unpolarized light is composed of many rays having random polarization directions.

- Light can be polarized by passing it through a polarizing filter or other polarizing material. The intensity of polarized light after passing through a polarizing filter is θ where is the original intensity and is the angle between the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Brewster's law states that reflected light will be completely polarized at the angle of reflection , known as Brewster's angle, given by a statement known as Brewster's law: —, where is the medium in which the incident and reflected light travel and is the index of refraction of the medium that forms the interface that reflects the light.
- Polarization can also be produced by scattering.
- There are a number of types of optically active substances that rotate the direction of polarization of light passing through them.

Conceptual Questions

Exercise:

Problem:

Under what circumstances is the phase of light changed by reflection? Is the phase related to polarization?

Exercise:

Problem: Can a sound wave in air be polarized? Explain.

Exercise:

Problem:

No light passes through two perfect polarizing filters with perpendicular axes. However, if a third polarizing filter is placed between the original two, some light can pass. Why is this? Under what circumstances does most of the light pass?

Exercise:

Problem:

Explain what happens to the energy carried by light that it is dimmed by passing it through two crossed polarizing filters.

Exercise:

Problem:

When particles scattering light are much smaller than its wavelength, the amount of scattering is proportional to . Does this mean there is more scattering for small than large ? How does this relate to the fact that the sky is blue?

Exercise:

Problem:

Using the information given in the preceding question, explain why sunsets are red.

Exercise:

Problem:

When light is reflected at Brewster's angle from a smooth surface, it is polarized parallel to the surface. Part of the light will be refracted into the surface. Describe how you would do an experiment to determine the polarization of the refracted light. What direction would you expect the polarization to have and would you expect it to be ?

Problems & Exercises

Exercise:

Problem:

What angle is needed between the direction of polarized light and the axis of a polarizing filter to cut its intensity in half?

Solution:

º

Exercise:

Problem:

The angle between the axes of two polarizing filters is º. By how much does the second filter reduce the intensity of the light coming through the first?

Exercise:

Problem:

If you have completely polarized light of intensity , what will its intensity be after passing through a polarizing filter with its axis at an º angle to the light's polarization direction?

Solution:

Exercise:

Problem:

What angle would the axis of a polarizing filter need to make with the direction of polarized light of intensity to reduce the intensity to ?

Exercise:

Problem:

At the end of [link], it was stated that the intensity of polarized light is reduced to of its original value by passing through a polarizing filter with its axis at an angle of º to the direction of polarization. Verify this statement.

Solution:

Exercise:

Problem:

Show that if you have three polarizing filters, with the second at an angle of º to the first and the third at an angle of º to the first, the intensity of light passed by the first will be reduced to of its value. (This is in contrast to having only the first and third, which reduces the intensity to zero, so that placing the second between them increases the intensity of the transmitted light.)

Exercise:

Problem:

Prove that, if is the intensity of light transmitted by two polarizing filters with axes at an angle and is the intensity when the axes are at an angle º θ then the original intensity. (Hint: Use the trigonometric identities º and)

Solution:

Exercise:

Problem:

At what angle will light reflected from diamond be completely polarized?

Exercise:

Problem:

What is Brewster's angle for light traveling in water that is reflected from crown glass?

Solution:

º

Exercise:

Problem:

A scuba diver sees light reflected from the water's surface. At what angle will this light be completely polarized?

Exercise:

Problem:

At what angle is light inside crown glass completely polarized when reflected from water, as in a fish tank?

Solution:

º

Exercise:

Problem:

Light reflected at º from a window is completely polarized. What is the window's index of refraction and the likely substance of which it is made?

Exercise:

Problem:

(a) Light reflected at º from a gemstone in a ring is completely polarized. Can the gem be a diamond? (b) At what angle would the light be completely polarized if the gem was in water?

Solution:

(a) 1.92, not diamond (Zircon)

(b) º

Exercise:

Problem:

If is Brewster's angle for light reflected from the top of an interface between two substances, and is Brewster's angle for light reflected from below, prove that º

Exercise:

Problem: Integrated Concepts

If a polarizing filter reduces the intensity of polarized light to of its original value, by how much are the electric and magnetic fields reduced?

Solution:

Exercise:

Problem: Integrated Concepts

Suppose you put on two pairs of Polaroid sunglasses with their axes at an angle of º. How much longer will it take the light to deposit a given amount of energy in your eye compared with a single pair of sunglasses? Assume the lenses are clear except for their polarizing characteristics.

Exercise:

Problem: Integrated Concepts

(a) On a day when the intensity of sunlight is , a circular lens 0.200 m in diameter focuses light onto water in a black beaker. Two polarizing sheets of plastic are placed in front of the lens with their axes at an angle of º Assuming the sunlight is unpolarized and the polarizers are efficient, what is the initial rate of heating of the water in º , assuming it is absorbed? The aluminum

beaker has a mass of 30.0 grams and contains 250 grams of water. (b) Do the polarizing filters get hot? Explain.

Solution:

(a) °C/s

(b) Yes, the polarizing filters get hot because they absorb some of the lost energy from the sunlight.

Glossary

axis of a polarizing filter
: the direction along which the filter passes the electric field of an EM wave

birefringent
: crystals that split an unpolarized beam of light into two beams

Brewster’s angle
: — where is the index of refraction of the medium from which the light is reflected and is the index of refraction of the medium in which the reflected light travels

Brewster’s law
: —, where is the medium in which the incident and reflected light travel and is the index of refraction of the medium that forms the interface that reflects the light

direction of polarization
: the direction parallel to the electric field for EM waves

horizontally polarized
: the oscillations are in a horizontal plane

optically active

substances that rotate the plane of polarization of light passing through them

polarization
: the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized
: waves having the electric and magnetic field oscillations in a definite direction

reflected light that is completely polarized
: light reflected at the angle of reflection , known as Brewster's angle

unpolarized
: waves that are randomly polarized

vertically polarized
: the oscillations are in a vertical plane

Extended Topic Microscopy Enhanced by the Wave Characteristics of Light

- Discuss the different types of microscopes.

Physics research underpins the advancement of developments in microscopy. As we gain knowledge of the wave nature of electromagnetic waves and methods to analyze and interpret signals, new microscopes that enable us to “see” more are being developed. It is the evolution and newer generation of microscopes that are described in this section.

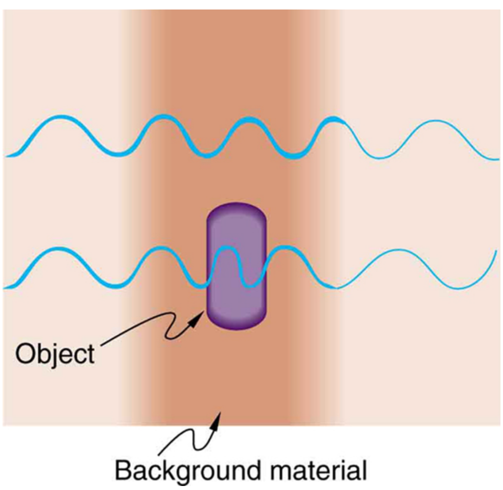
The use of microscopes (microscopy) to observe small details is limited by the wave nature of light. Owing to the fact that light diffracts significantly around small objects, it becomes impossible to observe details significantly smaller than the wavelength of light. One rule of thumb has it that all details smaller than about λ are difficult to observe. Radar, for example, can detect the size of an aircraft, but not its individual rivets, since the wavelength of most radar is several centimeters or greater. Similarly, visible light cannot detect individual atoms, since atoms are about 0.1 nm in size and visible wavelengths range from 380 to 760 nm. Ironically, special techniques used to obtain the best possible resolution with microscopes take advantage of the same wave characteristics of light that ultimately limit the detail.

Note:
Making Connections: Waves
All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Sonar and medical ultrasound are limited by the wavelength of sound they employ. We shall see that this is also true in electron microscopy, since electrons have a wavelength. Heisenberg’s uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in quantum mechanics.

The most obvious method of obtaining better detail is to utilize shorter wavelengths. **Ultraviolet (UV) microscopes** have been constructed with

special lenses that transmit UV rays and utilize photographic or electronic techniques to record images. The shorter UV wavelengths allow somewhat greater detail to be observed, but drawbacks, such as the hazard of UV to living tissue and the need for special detection devices and lenses (which tend to be dispersive in the UV), severely limit the use of UV microscopes. Elsewhere, we will explore practical uses of very short wavelength EM waves, such as x rays, and other short-wavelength probes, such as electrons in electron microscopes, to detect small details.

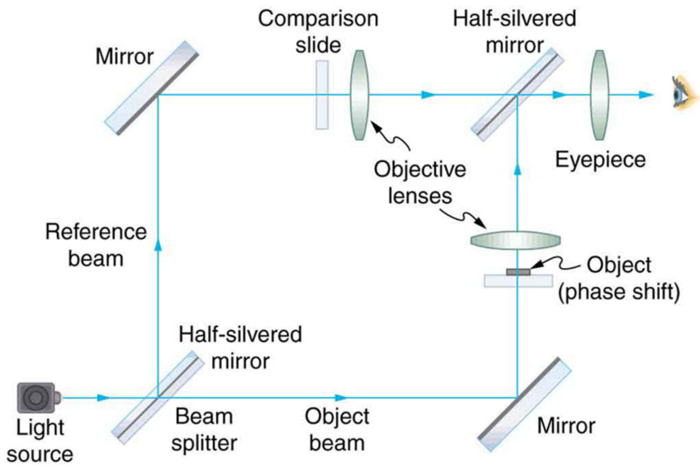
Another difficulty in microscopy is the fact that many microscopic objects do not absorb much of the light passing through them. The lack of contrast makes image interpretation very difficult. **Contrast** is the difference in intensity between objects and the background on which they are observed. Stains (such as dyes, fluorophores, etc.) are commonly employed to enhance contrast, but these tend to be application specific. More general wave interference techniques can be used to produce contrast. [link] shows the passage of light through a sample. Since the indices of refraction differ, the number of wavelengths in the paths differs. Light emerging from the object is thus out of phase with light from the background and will interfere differently, producing enhanced contrast, especially if the light is coherent and monochromatic—as in laser light.



Light rays passing through a sample under a microscope will emerge with different phases depending on their paths.

The object shown has a greater index of refraction than the background, and so the wavelength decreases as the ray passes through it. Superimposing these rays produces interference that varies with path, enhancing contrast between the object and background.

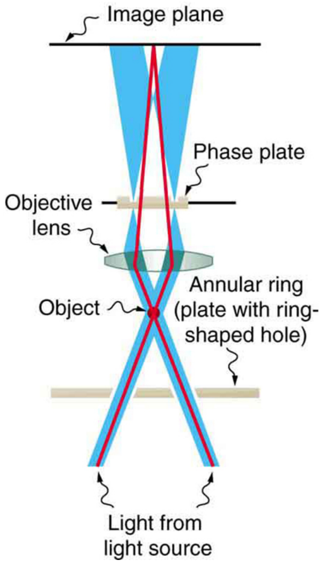
Interference microscopes enhance contrast between objects and background by superimposing a reference beam of light upon the light emerging from the sample. Since light from the background and objects differ in phase, there will be different amounts of constructive and destructive interference, producing the desired contrast in final intensity. [link] shows schematically how this is done. Parallel rays of light from a source are split into two beams by a half-silvered mirror. These beams are called the object and reference beams. Each beam passes through identical optical elements, except that the object beam passes through the object we wish to observe microscopically. The light beams are recombined by another half-silvered mirror and interfere. Since the light rays passing through different parts of the object have different phases, interference will be significantly different and, hence, have greater contrast between them.



An interference microscope utilizes interference between the reference and object beam to enhance contrast. The two beams are split by a half-silvered mirror; the object beam is sent through the object, and the reference beam is sent through otherwise identical optical elements. The beams are recombined by another half-silvered mirror, and the interference depends on the various phases emerging from different parts of the object, enhancing contrast.

Another type of microscope utilizing wave interference and differences in phases to enhance contrast is called the **phase-contrast microscope**. While its principle is the same as the interference microscope, the phase-contrast microscope is simpler to use and construct. Its impact (and the principle upon which it is based) was so important that its developer, the Dutch physicist Frits Zernike (1888–1966), was awarded the Nobel Prize in 1953. [link] shows the basic construction of a phase-contrast microscope. Phase differences between light passing through the object and background are produced by passing the rays through different parts of a phase plate (so called because it shifts the phase of the light passing through it). These two

light rays are superimposed in the image plane, producing contrast due to their interference.



Simplified construction of a phase-contrast microscope. Phase differences between light passing through the object and background are produced by passing the rays through different parts of a phase plate. The light rays are superimposed in the image plane,

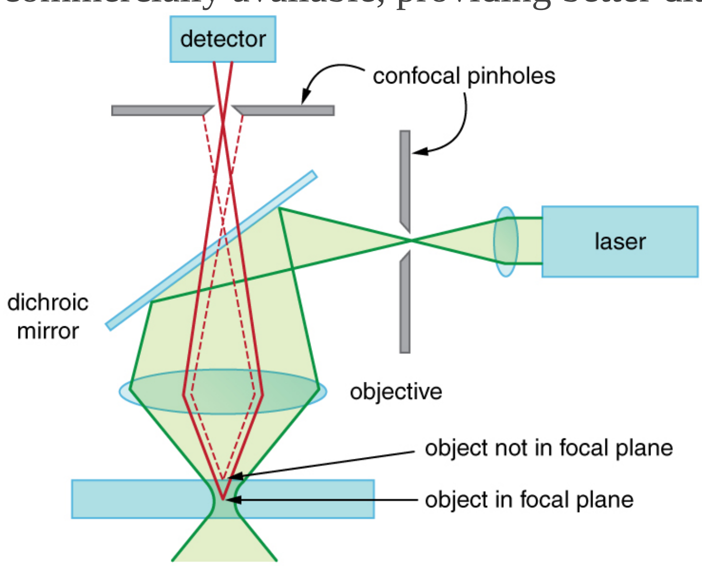
producing contrast due to their interference.

A **polarization microscope** also enhances contrast by utilizing a wave characteristic of light. Polarization microscopes are useful for objects that are optically active or birefringent, particularly if those characteristics vary from place to place in the object. Polarized light is sent through the object and then observed through a polarizing filter that is perpendicular to the original polarization direction. Nearly transparent objects can then appear with strong color and in high contrast. Many polarization effects are wavelength dependent, producing color in the processed image. Contrast results from the action of the polarizing filter in passing only components parallel to its axis.

Apart from the UV microscope, the variations of microscopy discussed so far in this section are available as attachments to fairly standard microscopes or as slight variations. The next level of sophistication is provided by commercial **confocal microscopes**, which use the extended focal region shown in [link](b) to obtain three-dimensional images rather than two-dimensional images. Here, only a single plane or region of focus is identified; out-of-focus regions above and below this plane are subtracted out by a computer so the image quality is much better. This type of microscope makes use of fluorescence, where a laser provides the excitation light. Laser light passing through a tiny aperture called a pinhole forms an extended focal region within the specimen. The reflected light passes through the objective lens to a second pinhole and the photomultiplier detector, see [link]. The second pinhole is the key here and serves to block much of the light from points that are not at the focal point of the objective lens. The pinhole is conjugate (coupled) to the focal point of the lens. The second pinhole and detector are scanned, allowing reflected light from a small region or section of the extended focal region to be imaged at any one time. The out-of-focus light is excluded. Each image is stored in a computer, and a full scanned image is generated in a short time. Live cell

processes can also be imaged at adequate scanning speeds allowing the imaging of three-dimensional microscopic movement. Confocal microscopy enhances images over conventional optical microscopy, especially for thicker specimens, and so has become quite popular.

The next level of sophistication is provided by microscopes attached to instruments that isolate and detect only a small wavelength band of light—monochromators and spectral analyzers. Here, the monochromatic light from a laser is scattered from the specimen. This scattered light shifts up or down as it excites particular energy levels in the sample. The uniqueness of the observed scattered light can give detailed information about the chemical composition of a given spot on the sample with high contrast—like molecular fingerprints. Applications are in materials science, nanotechnology, and the biomedical field. Fine details in biochemical processes over time can even be detected. The ultimate in microscopy is the electron microscope—to be discussed later. Research is being conducted into the development of new prototype microscopes that can become commercially available, providing better diagnostic and research capacities.



A confocal microscope provides three-dimensional images using pinholes and the extended depth of focus as described by wave optics. The right pinhole illuminates a tiny region of the sample in the focal

plane. In-focus light rays from this tiny region pass through the dichroic mirror and the second pinhole to a detector and a computer. Out-of-focus light rays are blocked. The pinhole is scanned sideways to form an image of the entire focal plane. The pinhole can then be scanned up and down to gather images from different focal planes. The result is a three-dimensional image of the specimen.

Section Summary

- To improve microscope images, various techniques utilizing the wave characteristics of light have been developed. Many of these enhance contrast with interference effects.

Conceptual Questions

Exercise:

Problem:

Explain how microscopes can use wave optics to improve contrast and why this is important.

Exercise:

Problem:

A bright white light under water is collimated and directed upon a prism. What range of colors does one see emerging?

Glossary

confocal microscopes
: microscopes that use the extended focal region to obtain three-dimensional images rather than two-dimensional images

contrast
: the difference in intensity between objects and the background on which they are observed

interference microscopes
: microscopes that enhance contrast between objects and background by superimposing a reference beam of light upon the light emerging from the sample

phase-contrast microscope
: microscope utilizing wave interference and differences in phases to enhance contrast

polarization microscope
: microscope that enhances contrast by utilizing a wave characteristic of light, useful for objects that are optically active

ultraviolet (UV) microscopes
: microscopes constructed with special lenses that transmit UV rays and utilize photographic or electronic techniques to record images

Atomic Masses

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
0	neutron	1	n	1.008 665	β^-	10.37 min
1	Hydrogen	1	^1H	1.007 825	99.985%	
	Deuterium	2	^2H or D	2.014 102	0.015%	
	Tritium	3	^3H or T	3.016 050	β^-	12.33 y
2	Helium	3	^3He	3.016 030	$1.38 \times 10^{-4}\%$	
		4	^4He	4.002 603	$\approx 100\%$	
3	Lithium	6	^6Li	6.015 121	7.5%	
		7	^7Li	7.016 003	92.5%	
4	Beryllium	7	^7Be	7.016 928	EC	53.29 d
		9	^9Be	9.012 182	100%	
5	Boron	10	^{10}B	10.012 937	19.9%	
		11	^{11}B	11.009 305	80.1%	
6	Carbon	11	^{11}C	11.011 432	EC, β^+	
		12	^{12}C	12.000 000	98.90%	
		13	^{13}C	13.003 355	1.10%	

Atomic Number, *Z*	**Name**	**Atomic Mass Number, *A***	**Symbol**	**Atomic Mass (u)**	**Percent Abundance or Decay Mode**	**Half-life, $t_{1/2}$**
		14	^{14}C	14.003 241	β^-	5730 y
7	Nitrogen	13	^{13}N	13.005 738	β^+	9.96 min
		14	^{14}N	14.003 074	99.63%	
		15	^{15}N	15.000 108	0.37%	
8	Oxygen	15	^{15}O	15.003 065	EC, β^+	122 s
		16	^{16}O	15.994 915	99.76%	
		18	^{18}O	17.999 160	0.200%	
9	Fluorine	18	^{18}F	18.000 937	EC, β^+	1.83 h
		19	^{19}F	18.998 403	100%	
10	Neon	20	^{20}Ne	19.992 435	90.51%	
		22	^{22}Ne	21.991 383	9.22%	
11	Sodium	22	^{22}Na	21.994 434	β^+	2.602 y
		23	^{23}Na	22.989 767	100%	
		24	^{24}Na	23.990 961	β^-	14.96 h
12	Magnesium	24	^{24}Mg	23.985 042	78.99%	
13	Aluminum	27	^{27}Al	26.981 539	100%	
14	Silicon	28	^{28}Si	27.976 927	92.23%	2.62h

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		31	^{31}Si	30.975 362	β^-	
15	Phosphorus	31	^{31}P	30.973 762	100%	
		32	^{32}P	31.973 907	β^-	14.28 d
16	Sulfur	32	^{32}S	31.972 070	95.02%	
		35	^{35}S	34.969 031	β^-	87.4 d
17	Chlorine	35	^{35}Cl	34.968 852	75.77%	
		37	^{37}Cl	36.965 903	24.23%	
18	Argon	40	^{40}Ar	39.962 384	99.60%	
19	Potassium	39	^{39}K	38.963 707	93.26%	
		40	^{40}K	39.963 999	0.0117%, EC, β^-	$1.28 \times 10^9\text{y}$
20	Calcium	40	^{40}Ca	39.962 591	96.94%	
21	Scandium	45	^{45}Sc	44.955 910	100%	
22	Titanium	48	^{48}Ti	47.947 947	73.8%	
23	Vanadium	51	^{51}V	50.943 962	99.75%	
24	Chromium	52	^{52}Cr	51.940 509	83.79%	
25	Manganese	55	^{55}Mn	54.938 047	100%	
26	Iron	56	^{56}Fe	55.934 939	91.72%	

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
27	Cobalt	59	^{59}Co	58.933 198	100%	
		60	^{60}Co	59.933 819	β^-	5.271 y
28	Nickel	58	^{58}Ni	57.935 346	68.27%	
		60	^{60}Ni	59.930 788	26.10%	
29	Copper	63	^{63}Cu	62.939 598	69.17%	
		65	^{65}Cu	64.927 793	30.83%	
30	Zinc	64	^{64}Zn	63.929 145	48.6%	
		66	^{66}Zn	65.926 034	27.9%	
31	Gallium	69	^{69}Ga	68.925 580	60.1%	
32	Germanium	72	^{72}Ge	71.922 079	27.4%	
		74	^{74}Ge	73.921 177	36.5%	
33	Arsenic	75	^{75}As	74.921 594	100%	
34	Selenium	80	^{80}Se	79.916 520	49.7%	
35	Bromine	79	^{79}Br	78.918 336	50.69%	
36	Krypton	84	^{84}Kr	83.911 507	57.0%	
37	Rubidium	85	^{85}Rb	84.911 794	72.17%	
38	Strontium	86	^{86}Sr	85.909 267	9.86%	

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		88	^{88}Sr	87.905 619	82.58%	
		90	^{90}Sr	89.907 738	β^-	28.8 y
39	Yttrium	89	^{89}Y	88.905 849	100%	
		90	^{90}Y	89.907 152	β^-	64.1 h
40	Zirconium	90	^{90}Zr	89.904 703	51.45%	
41	Niobium	93	^{93}Nb	92.906 377	100%	
42	Molybdenum	98	^{98}Mo	97.905 406	24.13%	
43	Technetium	98	^{98}Tc	97.907 215	β^-	$4.2 \times 10^6 y$
44	Ruthenium	102	^{102}Ru	101.904 348	31.6%	
45	Rhodium	103	^{103}Rh	102.905 500	100%	
46	Palladium	106	^{106}Pd	105.903 478	27.33%	
47	Silver	107	^{107}Ag	106.905 092	51.84%	
		109	^{109}Ag	108.904 757	48.16%	
48	Cadmium	114	^{114}Cd	113.903 357	28.73%	
49	Indium	115	^{115}In	114.903 880	95.7%, β^-	$4.4 \times 10^{14} y$
50	Tin	120	^{120}Sn	119.902 200	32.59%	
51	Antimony	121	^{121}Sb	120.903 821	57.3%	

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
52	Tellurium	130	^{130}Te	129.906 229	33.8%, β^-	$2.5 \times 10^{21}\text{y}$
53	Iodine	127	^{127}I	126.904 473	100%	
		131	^{131}I	130.906 114	β^-	8.040 d
54	Xenon	132	^{132}Xe	131.904 144	26.9%	
		136	^{136}Xe	135.907 214	8.9%	
55	Cesium	133	^{133}Cs	132.905 429	100%	
		134	^{134}Cs	133.906 696	EC, β^-	2.06 y
56	Barium	137	^{137}Ba	136.905 812	11.23%	
		138	^{138}Ba	137.905 232	71.70%	
57	Lanthanum	139	^{139}La	138.906 346	99.91%	
58	Cerium	140	^{140}Ce	139.905 433	88.48%	
59	Praseodymium	141	^{141}Pr	140.907 647	100%	
60	Neodymium	142	^{142}Nd	141.907 719	27.13%	
61	Promethium	145	^{145}Pm	144.912 743	EC, α	17.7 y
62	Samarium	152	^{152}Sm	151.919 729	26.7%	
63	Europium	153	^{153}Eu	152.921 225	52.2%	
64	Gadolinium	158	^{158}Gd	157.924 099	24.84%	

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
65	Terbium	159	^{159}Tb	158.925 342	100%	
66	Dysprosium	164	^{164}Dy	163.929 171	28.2%	
67	Holmium	165	^{165}Ho	164.930 319	100%	
68	Erbium	166	^{166}Er	165.930 290	33.6%	
69	Thulium	169	^{169}Tm	168.934 212	100%	
70	Ytterbium	174	^{174}Yb	173.938 859	31.8%	
71	Lutecium	175	^{175}Lu	174.940 770	97.41%	
72	Hafnium	180	^{180}Hf	179.946 545	35.10%	
73	Tantalum	181	^{181}Ta	180.947 992	99.98%	
74	Tungsten	184	^{184}W	183.950 928	30.67%	
75	Rhenium	187	^{187}Re	186.955 744	62.6%, β^-	$4.6 \times 10^{10}\text{y}$
76	Osmium	191	^{191}Os	190.960 920	β^-	15.4 d
		192	^{192}Os	191.961 467	41.0%	
77	Iridium	191	^{191}Ir	190.960 584	37.3%	
		193	^{193}Ir	192.962 917	62.7%	
78	Platinum	195	^{195}Pt	194.964 766	33.8%	
79	Gold	197	^{197}Au	196.966 543	100%	

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		198	$^{198}\mathrm{Au}$	197.968 217	β^-	2.696 d
80	Mercury	199	$^{199}\mathrm{Hg}$	198.968 253	16.87%	
		202	$^{202}\mathrm{Hg}$	201.970 617	29.86%	
81	Thallium	205	$^{205}\mathrm{Tl}$	204.974 401	70.48%	
82	Lead	206	$^{206}\mathrm{Pb}$	205.974 440	24.1%	
		207	$^{207}\mathrm{Pb}$	206.975 872	22.1%	
		208	$^{208}\mathrm{Pb}$	207.976 627	52.4%	
		210	$^{210}\mathrm{Pb}$	209.984 163	α, β^-	22.3 y
		211	$^{211}\mathrm{Pb}$	210.988 735	β^-	36.1 min
		212	$^{212}\mathrm{Pb}$	211.991 871	β^-	10.64 h
83	Bismuth	209	$^{209}\mathrm{Bi}$	208.980 374	100%	
		211	$^{211}\mathrm{Bi}$	210.987 255	α, β^-	2.14 min
84	Polonium	210	$^{210}\mathrm{Po}$	209.982 848	α	138.38 d
85	Astatine	218	$^{218}\mathrm{At}$	218.008 684	α, β^-	1.6 s
86	Radon	222	$^{222}\mathrm{Rn}$	222.017 570	α	3.82 d
87	Francium	223	$^{223}\mathrm{Fr}$	223.019 733	α, β^-	21.8 min
88	Radium	226	$^{226}\mathrm{Ra}$	226.025 402	α	1.60×10^3y

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
89	Actinium	227	^{227}Ac	227.027 750	α, β^-	21.8 y
90	Thorium	228	^{228}Th	228.028 715	α	1.91 y
		232	^{232}Th	232.038 054	100%, α	$1.41 \times 10^{10}\text{y}$
91	Protactinium	231	^{231}Pa	231.035 880	α	$3.28 \times 10^{4}\text{y}$
92	Uranium	233	^{233}U	233.039 628	α	$1.59 \times 10^{3}\text{y}$
		235	^{235}U	235.043 924	0.720%, α	$7.04 \times 10^{8}\text{y}$
		236	^{236}U	236.045 562	α	$2.34 \times 10^{7}\text{y}$
		238	^{238}U	238.050 784	99.2745%, α	$4.47 \times 10^{9}\text{y}$
		239	^{239}U	239.054 289	β^-	23.5 min
93	Neptunium	239	^{239}Np	239.052 933	β^-	2.355 d
94	Plutonium	239	^{239}Pu	239.052 157	α	$2.41 \times 10^{4}\text{y}$
95	Americium	243	^{243}Am	243.061 375	α, fission	$7.37 \times 10^{3}\text{y}$
96	Curium	245	^{245}Cm	245.065 483	α	$8.50 \times 10^{3}\text{y}$
97	Berkelium	247	^{247}Bk	247.070 300	α	$1.38 \times 10^{3}\text{y}$
98	Californium	249	^{249}Cf	249.074 844	α	351 y
99	Einsteinium	254	^{254}Es	254.088 019	α, β^-	276 d
100	Fermium	253	^{253}Fm	253.085 173	EC, α	3.00 d

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
101	Mendelevium	255	^{255}Md	255.091 081	EC, α	27 min
102	Nobelium	255	^{255}No	255.093 260	EC, α	3.1 min
103	Lawrencium	257	^{257}Lr	257.099 480	EC, α	0.646 s
104	Rutherfordium	261	^{261}Rf	261.108 690	α	1.08 min
105	Dubnium	262	^{262}Db	262.113 760	α, fission	34 s
106	Seaborgium	263	^{263}Sg	263.11 86	α, fission	0.8 s
107	Bohrium	262	^{262}Bh	262.123 1	α	0.102 s
108	Hassium	264	^{264}Hs	264.128 5	α	0.08 ms
109	Meitnerium	266	^{266}Mt	266.137 8	α	3.4 ms

Atomic Masses

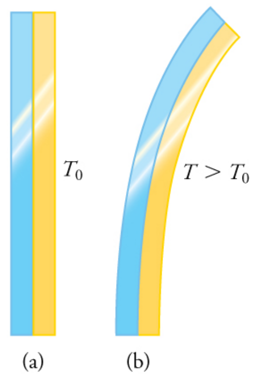
Temperature

- Define temperature.
- Convert temperatures between the Celsius, Fahrenheit, and Kelvin scales.
- Define thermal equilibrium.
- State the zeroth law of thermodynamics.

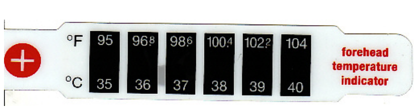
The concept of temperature has evolved from the common concepts of hot and cold. Human perception of what feels hot or cold is a relative one. For example, if you place one hand in hot water and the other in cold water, and then place both hands in tepid water, the tepid water will feel cool to the hand that was in hot water, and warm to the one that was in cold water. The scientific definition of temperature is less ambiguous than your senses of hot and cold. **Temperature** is operationally defined to be what we measure with a thermometer. (Many physical quantities are defined solely in terms of how they are measured. We shall see later how temperature is related to the kinetic energies of atoms and molecules, a more physical explanation.) Two accurate thermometers, one placed in hot water and the other in cold water, will show the hot water to have a higher temperature. If they are then placed in the tepid water, both will give identical readings (within measurement uncertainties). In this section, we discuss temperature, its measurement by thermometers, and its relationship to thermal equilibrium. Again, temperature is the quantity measured by a thermometer.

Note:
Misconception Alert: Human Perception vs. Reality
On a cold winter morning, the wood on a porch feels warmer than the metal of your bike. The wood and bicycle are in thermal equilibrium with the outside air, and are thus the same temperature. They *feel* different because of the difference in the way that they conduct heat away from your skin. The metal conducts heat away from your body faster than the wood does (see more about conductivity in Conduction). This is just one example demonstrating that the human sense of hot and cold is not determined by temperature alone.
Another factor that affects our perception of temperature is humidity. Most people feel much hotter on hot, humid days than on hot, dry days. This is because on humid days, sweat does not evaporate from the skin as efficiently as it does on dry days. It is the evaporation of sweat (or water from a sprinkler or pool) that cools us off.

Any physical property that depends on temperature, and whose response to temperature is reproducible, can be used as the basis of a thermometer. Because many physical properties depend on temperature, the variety of thermometers is remarkable. For example, volume increases with temperature for most substances. This property is the basis for the common alcohol thermometer, the old mercury thermometer, and the bimetallic strip ([link]). Other properties used to measure temperature include electrical resistance and color, as shown in [link], and the emission of infrared radiation, as shown in [link].



The curvature of a bimetallic strip depends on temperature. (a) The strip is straight at the starting temperature, where its two components have the same length. (b) At a higher temperature, this strip bends to the right, because the metal on the left has expanded more than the metal on the right.



Each of the six squares on this plastic (liquid crystal)

thermometer contains a film of a different heat-sensitive liquid crystal material. Below °, all six squares are black. When the plastic thermometer is exposed to temperature that increases to °, the first liquid crystal square changes color. When the temperature increases above ° the second liquid crystal square also changes color, and so forth. (credit: Arkrishna, Wikimedia Commons)

Fireman Jason Ormand uses a pyrometer to check the temperature of an aircraft carrier's ventilation system. Infrared radiation (whose emission varies with temperature)

from the vent is measured and a temperature readout is quickly produced. Infrared measurements are also frequently used as a measure of body temperature. These modern thermometers, placed in the ear canal, are more accurate than alcohol thermometers placed under the tongue or in the armpit. (credit: Lamel J. Hinton/U.S. Navy)

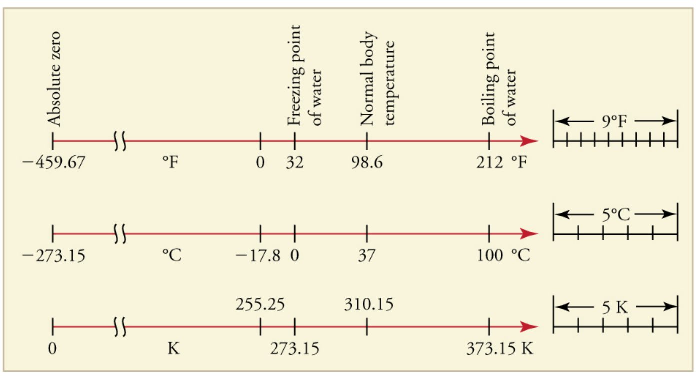
Temperature Scales

Thermometers are used to measure temperature according to well-defined scales of measurement, which use pre-defined reference points to help compare quantities. The three most common temperature scales are the Fahrenheit, Celsius, and Kelvin scales. A temperature scale can be created by identifying two easily reproducible temperatures. The freezing and boiling temperatures of water at standard atmospheric pressure are commonly used.

The **Celsius** scale (which replaced the slightly different *centigrade* scale) has the freezing point of water at $^{\circ}$ and the boiling point at $^{\circ}$. Its unit is the **degree Celsius** $^{\circ}$. On the **Fahrenheit** scale (still the most frequently used in the United States), the freezing point of water is at $^{\circ}$ and the boiling point is at $^{\circ}$. The unit of temperature on this scale is the **degree Fahrenheit** $^{\circ}$. Note that a temperature difference of one degree Celsius is greater than a temperature difference of one degree Fahrenheit. Only 100 Celsius degrees

span the same range as 180 Fahrenheit degrees, thus one degree on the Celsius scale is 1.8 times larger than one degree on the Fahrenheit scale

The **Kelvin** scale is the temperature scale that is commonly used in science. It is an *absolute temperature* scale defined to have 0 K at the lowest possible temperature, called **absolute zero**. The official temperature unit on this scale is the *kelvin*, which is abbreviated K, and is not accompanied by a degree sign. The freezing and boiling points of water are 273.15 K and 373.15 K, respectively. Thus, the magnitude of temperature differences is the same in units of kelvins and degrees Celsius. Unlike other temperature scales, the Kelvin scale is an absolute scale. It is used extensively in scientific work because a number of physical quantities, such as the volume of an ideal gas, are directly related to absolute temperature. The kelvin is the SI unit used in scientific work.



Relationships between the Fahrenheit, Celsius, and Kelvin temperature scales, rounded to the nearest degree. The relative sizes of the scales are also shown.

The relationships between the three common temperature scales is shown in [link]. Temperatures on these scales can be converted using the equations in [link].

To convert from . . .	Use this equation . . .	Also written as . . .

To convert from . . .	**Use this equation . . .**	**Also written as . . .**
Celsius to Fahrenheit	o — o	o — o
Fahrenheit to Celsius	o — o	o — o
Celsius to Kelvin	o	o
Kelvin to Celsius	o	o
Fahrenheit to Kelvin	— o	— o
Kelvin to Fahrenheit	o —	o —

Temperature Conversions

Notice that the conversions between Fahrenheit and Kelvin look quite complicated. In fact, they are simple combinations of the conversions between Fahrenheit and Celsius, and the conversions between Celsius and Kelvin.

Example:
Converting between Temperature Scales: Room Temperature
“Room temperature” is generally defined to be o . (a) What is room temperature in o ? (b) What is it in K?
Strategy
To answer these questions, all we need to do is choose the correct conversion equations and plug in the known values.

Solution for (a)

1. Choose the right equation. To convert from º to º , use the equation

Equation:

2. Plug the known value into the equation and solve:

Equation:

Solution for (b)

1. Choose the right equation. To convert from º to K, use the equation

Equation:

2. Plug the known value into the equation and solve:

Equation:

Example:

Converting between Temperature Scales: the Reaumur Scale

The Reaumur scale is a temperature scale that was used widely in Europe in the 18th and 19th centuries. On the Reaumur temperature scale, the freezing point of water is º and the boiling temperature is º . If "room temperature" is º on the Celsius scale, what is it on the Reaumur scale?

Strategy

To answer this question, we must compare the Reaumur scale to the Celsius scale. The difference between the freezing point and boiling point of water on the Reaumur scale is º . On the Celsius scale it is º . Therefore º º . Both scales start at º for freezing, so we can derive a simple formula to convert between temperatures on the two scales.

Solution

1. Derive a formula to convert from one scale to the other:

Equation:

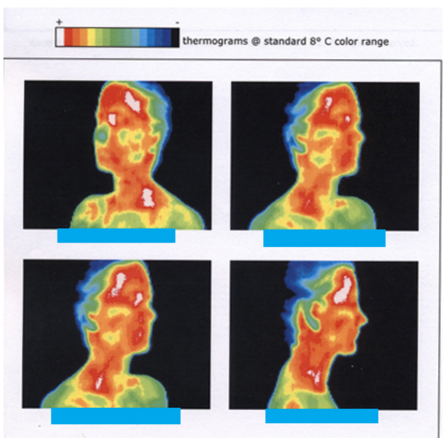
2. Plug the known value into the equation and solve:

Equation:

$$^{\circ} \quad \frac{^{\circ}}{^{\circ}} \quad ^{\circ} \quad ^{\circ}$$

Temperature Ranges in the Universe

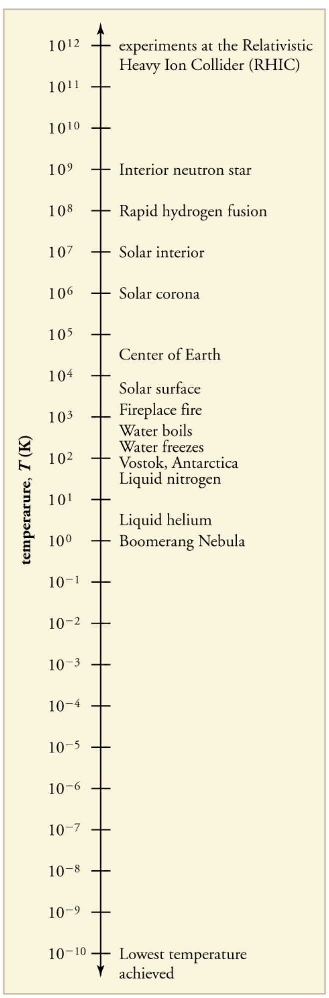
[link] shows the wide range of temperatures found in the universe. Human beings have been known to survive with body temperatures within a small range, from $^{\circ}$ to $^{\circ}$ $^{\circ}$ to $^{\circ}$). The average normal body temperature is usually given as $^{\circ}$ ($^{\circ}$), and variations in this temperature can indicate a medical condition: a fever, an infection, a tumor, or circulatory problems (see [link]).



This image of radiation from a person’s body (an infrared thermograph) shows the location of temperature abnormalities in the upper body. Dark blue corresponds to cold areas and red to white corresponds to hot areas. An elevated temperature might be an indication of malignant tissue (a cancerous tumor in the breast, for example), while a depressed temperature

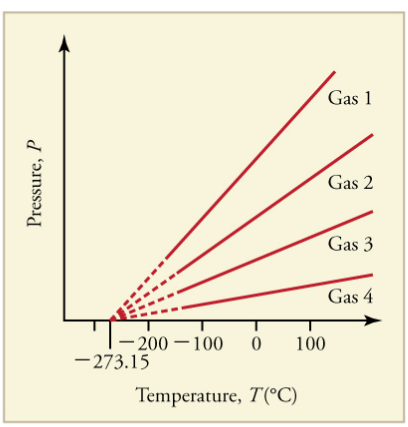
might be due to a decline in blood flow from a clot. In this case, the abnormalities are caused by a condition called hyperhidrosis. (credit: Porcelina81, Wikimedia Commons)

The lowest temperatures ever recorded have been measured during laboratory experiments: at the Massachusetts Institute of Technology (USA), and at Helsinki University of Technology (Finland). In comparison, the coldest recorded place on Earth's surface is Vostok, Antarctica at 183 K o , and the coldest place (outside the lab) known in the universe is the Boomerang Nebula, with a temperature of 1 K.



Each increment on this logarithmic scale indicates an increase by a factor of ten, and thus illustrates the tremendous range of temperatures in nature. Note that zero on a logarithmic scale would occur off the bottom of the page at infinity.

Note:
Making Connections: Absolute Zero
What is absolute zero? Absolute zero is the temperature at which all molecular motion has ceased. The concept of absolute zero arises from the behavior of gases. [link] shows how the pressure of gases at a constant volume decreases as temperature decreases. Various scientists have noted that the pressures of gases extrapolate to zero at the same temperature, º . This extrapolation implies that there is a lowest temperature. This temperature is called *absolute zero*. Today we know that most gases first liquefy and then freeze, and it is not actually possible to reach absolute zero. The numerical value of absolute zero temperature is º or 0 K.



Graph of pressure versus temperature for various

gases kept at a constant volume. Note that all of the graphs extrapolate to zero pressure at the same temperature.

Thermal Equilibrium and the Zeroth Law of Thermodynamics

Thermometers actually take their *own* temperature, not the temperature of the object they are measuring. This raises the question of how we can be certain that a thermometer measures the temperature of the object with which it is in contact. It is based on the fact that any two systems placed in *thermal contact* (meaning heat transfer can occur between them) will reach the same temperature. That is, heat will flow from the hotter object to the cooler one until they have exactly the same temperature. The objects are then in **thermal equilibrium**, and no further changes will occur. The systems interact and change because their temperatures differ, and the changes stop once their temperatures are the same. Thus, if enough time is allowed for this transfer of heat to run its course, the temperature a thermometer registers *does* represent the system with which it is in thermal equilibrium. Thermal equilibrium is established when two bodies are in contact with each other and can freely exchange energy.

Furthermore, experimentation has shown that if two systems, A and B, are in thermal equilibrium with each another, and B is in thermal equilibrium with a third system C, then A is also in thermal equilibrium with C. This conclusion may seem obvious, because all three have the same temperature, but it is basic to thermodynamics. It is called the **zeroth law of thermodynamics**.

Note:
The Zeroth Law of Thermodynamics
If two systems, A and B, are in thermal equilibrium with each other, and B is in thermal equilibrium with a third system, C, then A is also in thermal equilibrium with C.

This law was postulated in the 1930s, after the first and second laws of thermodynamics had been developed and named. It is called the *zeroth law* because it comes logically before the first and second laws (discussed in Thermodynamics). An example of this law in action is seen in babies in incubators: babies in incubators normally have very few clothes on, so to an observer they look as if they may not be warm enough. However, the temperature of the air, the cot, and the baby is the same, because they are in thermal equilibrium, which is accomplished by maintaining air temperature to keep the baby comfortable.

Exercise:
Check Your Understanding

Problem: Does the temperature of a body depend on its size?

Solution:

No, the system can be divided into smaller parts each of which is at the same temperature. We say that the temperature is an *intensive* quantity. Intensive quantities are independent of size.

Section Summary

- Temperature is the quantity measured by a thermometer.
- Temperature is related to the average kinetic energy of atoms and molecules in a system.
- Absolute zero is the temperature at which there is no molecular motion.
- There are three main temperature scales: Celsius, Fahrenheit, and Kelvin.
- Temperatures on one scale can be converted to temperatures on another scale using the following equations:

  Equation:

  $$^\circ \quad — \quad ^\circ$$

  Equation:

  $$^\circ \quad — \quad ^\circ$$

  Equation:

  $$^\circ$$

  Equation:

  $$^\circ$$

- Systems are in thermal equilibrium when they have the same temperature.
- Thermal equilibrium occurs when two bodies are in contact with each other and can freely exchange energy.
- The zeroth law of thermodynamics states that when two systems, A and B, are in thermal equilibrium with each other, and B is in thermal equilibrium with a third system, C, then A is also in thermal equilibrium with C.

Conceptual Questions

Exercise:

Problem: What does it mean to say that two systems are in thermal equilibrium?

Exercise:

Problem:

Give an example of a physical property that varies with temperature and describe how it is used to measure temperature.

Exercise:

Problem:

When a cold alcohol thermometer is placed in a hot liquid, the column of alcohol goes *down* slightly before going up. Explain why.

Exercise:

Problem:

If you add boiling water to a cup at room temperature, what would you expect the final equilibrium temperature of the unit to be? You will need to include the surroundings as part of the system. Consider the zeroth law of thermodynamics.

Problems & Exercises

Exercise:

Problem: What is the Fahrenheit temperature of a person with a º fever?

Solution:

º

Exercise:

Problem:

Frost damage to most plants occurs at temperatures of º or lower. What is this temperature on the Kelvin scale?

Exercise:

Problem:

To conserve energy, room temperatures are kept at º in the winter and º in the summer. What are these temperatures on the Celsius scale?

Solution:

º and º

Exercise:

Problem:

A tungsten light bulb filament may operate at 2900 K. What is its Fahrenheit temperature? What is this on the Celsius scale?

Exercise:

Problem:

The surface temperature of the Sun is about 5750 K. What is this temperature on the Fahrenheit scale?

Solution:

º

Exercise:

Problem:

One of the hottest temperatures ever recorded on the surface of Earth was º in Death Valley, CA. What is this temperature in Celsius degrees? What is this temperature in Kelvin?

Exercise:

Problem:

(a) Suppose a cold front blows into your locale and drops the temperature by 40.0 Fahrenheit degrees. How many degrees Celsius does the temperature decrease when there is a º decrease in temperature? (b) Show that any change in temperature in Fahrenheit degrees is nine-fifths the change in Celsius degrees.

Solution:

(a) º

$^{\circ}$ $^{\circ}$ $^{\circ}$

(b) $-$ $^{\circ}$ $^{\circ}$ $-$ $^{\circ}$ $^{\circ}$

$-$ $^{\circ}$ $^{\circ}$ $-$ $^{\circ}$

Exercise:

Problem:

(a) At what temperature do the Fahrenheit and Celsius scales have the same numerical value? (b) At what temperature do the Fahrenheit and Kelvin scales have the same numerical value?

Glossary

temperature
: the quantity measured by a thermometer

Celsius scale
: temperature scale in which the freezing point of water is $^{\circ}$ and the boiling point of water is $^{\circ}$

degree Celsius
: unit on the Celsius temperature scale

Fahrenheit scale
: temperature scale in which the freezing point of water is $^{\circ}$ and the boiling point of water is $^{\circ}$

degree Fahrenheit
: unit on the Fahrenheit temperature scale

Kelvin scale
: temperature scale in which 0 K is the lowest possible temperature, representing absolute zero

absolute zero
: the lowest possible temperature; the temperature at which all molecular motion ceases

thermal equilibrium
: the condition in which heat no longer flows between two objects that are in contact; the two objects have the same temperature

zeroth law of thermodynamics
: law that states that if two objects are in thermal equilibrium, and a third object is in thermal equilibrium with one of those objects, it is also in thermal equilibrium with the other object

Selected Radioactive Isotopes

Decay modes are α, β^-, β^+, electron capture (EC) and isomeric transition (IT). EC results in the same daughter nucleus as would β^+ decay. IT is a transition from a metastable excited state. Energies for $\beta^\pm$ decays are the maxima; average energies are roughly one-half the maxima.

Isotope	***t* 1/2**	**DecayMode(s)**	**Energy(MeV)**	**Percent**		**γ-Ray Energy(MeV)**
^{3}H	12.33 y	β^-	0.0186	100%		
^{14}C	5730 y	β^-	0.156	100%		
^{13}N	9.96 min	β^+	1.20	100%		
^{22}Na	2.602 y	β^+	0.55	90%	γ	1.27
^{32}P	14.28 d	β^-	1.71	100%		
^{35}S	87.4 d	β^-	0.167	100%		
^{36}Cl	3.00×10^5y	β^-	0.710	100%		
^{40}K	1.28×10^9y	β^-	1.31	89%		
^{43}K	22.3 h	β^-	0.827	87%	γs	0.373
						0.618
^{45}Ca	165 d	β^-	0.257	100%		
^{51}Cr	27.70 d	EC			γ	0.320
^{52}Mn	5.59d	β^+	3.69	28%	γs	1.33
						1.43
^{52}Fe	8.27 h	β^+	1.80	43%		0.169
						0.378
^{59}Fe	44.6 d	β^-s	0.273	45%	γs	1.10
			0.466	55%		1.29
^{60}Co	5.271 y	β^-	0.318	100%	γs	1.17
						1.33
^{65}Zn	244.1 d	EC			γ	1.12

Isotope	$t_{1/2}$	DecayMode(s)	Energy(MeV)	Percent		γ-Ray Energy(MeV)
^{67}Ga	78.3 h	EC			γs	0.0933
						0.185
						0.300
						others
^{75}Se	118.5 d	EC			γs	0.121
						0.136
						0.265
						0.280
						others
^{86}Rb	18.8 d	β^-s	0.69	9%	γ	1.08
			1.77	91%		
^{85}Sr	64.8 d	EC			γ	0.514
^{90}Sr	28.8 y	β^-	0.546	100%		
^{90}Y	64.1 h	β^-	2.28	100%		
^{99m}Tc	6.02 h	IT			γ	0.142
^{113m}In	99.5 min	IT			γ	0.392
^{123}I	13.0 h	EC			γ	0.159
^{131}I	8.040 d	β^-s	0.248	7%	γs	0.364
			0.607	93%		others
			others			
^{129}Cs	32.3 h	EC			γs	0.0400
						0.372
						0.411
						others
^{137}Cs	30.17 y	β^-s	0.511	95%	γ	0.662
			1.17	5%		

Isotope	$t_{1/2}$	DecayMode(s)	Energy(MeV)	Percent		γ-Ray Energy(MeV)
^{140}Ba	12.79 d	β^-	1.035	$\approx$100%	γs	0.030
						0.044
						0.537
						others
^{198}Au	2.696 d	β^-	1.161	$\approx$100%	γ	0.412
^{197}Hg	64.1 h	EC			γ	0.0733
^{210}Po	138.38 d	α	5.41	100%		
^{226}Ra	1.60×10^3y	αs	4.68	5%	γ	0.186
			4.87	95%		
^{235}U	7.038×10^8y	α	4.68	$\approx$100%	γs	numerous
^{238}U	4.468×10^9y	αs	4.22	23%	γ	0.050
			4.27	77%		
^{237}Np	2.14×10^6y	αs	numerous		γs	numerous
			4.96 (max.)			
^{239}Pu	2.41×10^4y	αs	5.19	11%	γs	7.5×10^{-5}
			5.23	15%		0.013
			5.24	73%		0.052
						others
^{243}Am	7.37×10^3y	αs	Max. 5.44		γs	0.075
			5.37	88%		others
			5.32	11%		
			others			

Selected Radioactive Isotopes

Useful Information

This appendix is broken into several tables.

- [link], Important Constants
- [link], Submicroscopic Masses
- [link], Solar System Data
- [link], Metric Prefixes for Powers of Ten and Their Symbols
- [link], The Greek Alphabet
- [link], SI units
- [link], Selected British Units
- [link], Other Units
- [link], Useful Formulae

Symbol	Meaning	Best Value	Approximate Value
	Speed of light in vacuum		
	Gravitational constant		
	Avogadro's number		
	Boltzmann's constant		
	Gas constant		
σ	Stefan-Boltzmann constant		
	Coulomb force constant		
	Charge on electron		
ε	Permittivity of free space		
μ	Permeability of free space	π	
	Planck's constant		

Important Constants[footnote]

Stated values are according to the National Institute of Standards and Technology Reference on Constants, Units, an
www.physics.nist.gov/cuu (accessed May 18, 2012). Values in parentheses are the uncertainties in the last digits. Nu
are exact as defined.

Symbol	Meaning	Best Value	Approximate Value
	Electron mass		
	Proton mass		
	Neutron mass		
	Atomic mass unit		

Submicroscopic Masses[footnote]
Stated values are according to the National Institute of Standards and Technology Reference on Constants, Units, and Uncertainty, www.physics.nist.gov/cuu (accessed May 18, 2012). Values in parentheses are the uncertainties in the last digits. Numbers without uncertainties are exact as defined.

Sun	mass	
	average radius	
	Earth-sun distance (average)	
Earth	mass	
	average radius	
	orbital period	

Moon	mass	
	average radius	
	orbital period (average)	
	Earth-moon distance (average)	

Solar System Data

Prefix	**Symbol**	**Value**	**Prefix**	**Symbol**	**Value**
tera	T		deci	d	
giga	G		centi	c	
mega	M		milli	m	
kilo	k		micro		
hecto	h		nano	n	
deka	da		pico	p	
—	—		femto	f	

Metric Prefixes for Powers of Ten and Their Symbols

Alpha			Eta			Nu			Tau		
Beta			Theta			Xi			Upsilon		
Gamma			Iota			Omicron			Phi		
Delta			Kappa			Pi			Chi		

Epsilon			Lambda			Rho			Psi		
Zeta			Mu			Sigma			Omega		

The Greek Alphabet

	Entity	**Abbreviation**	**Name**
Fundamental units	Length	m	meter
	Mass	kg	kilogram
	Time	s	second
	Current	A	ampere
Supplementary unit	Angle	rad	radian
Derived units	Force		newton
	Energy		joule
	Power		watt
	Pressure		pascal
	Frequency		hertz
	Electronic potential		volt
	Capacitance		farad
	Charge		coulomb
	Resistance		ohm

	Entity	Abbreviation	Name
	Magnetic field		tesla
	Nuclear decay rate		becquerel

SI Units

Length	
Force	
Energy	
Power	
Pressure	

Selected British Units

Length	
	Å
Area	
Volume	

Mass	
Time	
Speed	
Angle	
Energy	
Pressure	
Nuclear decay rate	

Other Units

Circumference of a circle with radius or diameter	
Area of a circle with radius or diameter	
Area of a sphere with radius	

Volume of a sphere with radius	

Useful Formulae

Glossary of Key Symbols and Notation

In this glossary, key symbols and notation are briefly defined.

Symbol	Definition
any symbol	average (indicated by a bar over a symbol—e.g., $\bar{v}$ is average velocity)
$^\circ\text{C}$	Celsius degree
$^\circ\text{F}$	Fahrenheit degree
$//$	parallel
$\perp$	perpendicular
$\propto$	proportional to
$\pm$	plus or minus

Symbol	**Definition**
0	zero as a subscript denotes an initial value
α	alpha rays
α	angular acceleration
α	temperature coefficient(s) of resistivity
β	beta rays
β	sound level
β	volume coefficient of expansion
β^-	electron emitted in nuclear beta decay
β^+	positron decay
γ	gamma rays

Symbol	Definition
γ	surface tension
$\gamma = 1/\sqrt{1 - v^2/c^2}$	a constant used in relativity
Δ	change in whatever quantity follows
δ	uncertainty in whatever quantity follows
$\mathit{\Delta} E$	change in energy between the initial and final orbits of an electron in an atom
$\mathit{\Delta} E$	uncertainty in energy
$\mathit{\Delta} m$	difference in mass between initial and final products
$\mathit{\Delta} N$	number of decays that occur
$\mathit{\Delta} p$	change in momentum

Symbol	Definition
Δp	uncertainty in momentum
$\Delta \mathrm{PE_g}$	change in gravitational potential energy
$\Delta\theta$	rotation angle
Δs	distance traveled along a circular path
Δt	uncertainty in time
Δt_0	proper time as measured by an observer at rest relative to the process
ΔV	potential difference
Δx	uncertainty in position
ε_0	permittivity of free space
η	viscosity

Symbol	Definition
θ	angle between the force vector and the displacement vector
θ	angle between two lines
θ	contact angle
θ	direction of the resultant
θ_b	Brewster's angle
θ_c	critical angle
κ	dielectric constant
λ	decay constant of a nuclide
λ	wavelength
λ_n	wavelength in a medium

Symbol	Definition
μ_0	permeability of free space
μ_{k}	coefficient of kinetic friction
μ_{s}	coefficient of static friction
v_e	electron neutrino
π^+	positive pion
π^-	negative pion
π^0	neutral pion
ρ	density
ρ_{c}	critical density, the density needed to just halt universal expansion
ρ_{fl}	fluid density

Symbol	Definition
ρ_{obj}	average density of an object
ρ/ρ_{w}	specific gravity
τ	characteristic time constant for a resistance and inductance (RL) or resistance and capacitance (RC) circuit
τ	characteristic time for a resistor and capacitor (RC) circuit
τ	torque
Υ	upsilon meson
Φ	magnetic flux
ϕ	phase angle
Ω	ohm (unit)
ω	angular velocity

Symbol	Definition
A	ampere (current unit)
A	area
A	cross-sectional area
A	total number of nucleons
a	acceleration
a_{B}	Bohr radius
a_{c}	centripetal acceleration
a_{t}	tangential acceleration
AC	alternating current
AM	amplitude modulation

Symbol	Definition
atm	atmosphere
B	baryon number
B	blue quark color
B	antiblue (yellow) antiquark color
b	quark flavor bottom or beauty
B	bulk modulus
B	magnetic field strength
$\mathrm{B}_{\mathrm{int}}$	electron’s intrinsic magnetic field
$\mathrm{B}_{\mathrm{orb}}$	orbital magnetic field
BE	binding energy of a nucleus—it is the energy required to completely disassemble it into separate protons and neutrons

Symbol	Definition
BE/A	binding energy per nucleon
Bq	becquerel—one decay per second
C	capacitance (amount of charge stored per volt)
C	coulomb (a fundamental SI unit of charge)
C_{p}	total capacitance in parallel
C_{s}	total capacitance in series
CG	center of gravity
CM	center of mass
c	quark flavor charm
c	specific heat

Symbol	Definition
c	speed of light
Cal	kilocalorie
cal	calorie
COP_{hp}	heat pump’s coefficient of performance
COP_{ref}	coefficient of performance for refrigerators and air conditioners
$\cos\theta$	cosine
$\cot\theta$	cotangent
$\csc\theta$	cosecant
D	diffusion constant
d	displacement

Symbol	Definition
d	quark flavor down
dB	decibel
d_{i}	distance of an image from the center of a lens
d_{o}	distance of an object from the center of a lens
DC	direct current
E	electric field strength
ε	emf (voltage) or Hall electromotive force
emf	electromotive force
E	energy of a single photon
E	nuclear reaction energy

Symbol	Definition
E	relativistic total energy
E	total energy
E_0	ground state energy for hydrogen
E_0	rest energy
EC	electron capture
E_{cap}	energy stored in a capacitor
Eff	efficiency—the useful work output divided by the energy input
$\mathrm{Eff}_{\mathrm{C}}$	Carnot efficiency
E_{in}	energy consumed (food digested in humans)
E_{ind}	energy stored in an inductor

Symbol	Definition
E_{out}	energy output
e	emissivity of an object
e^{+}	antielectron or positron
eV	electron volt
F	farad (unit of capacitance, a coulomb per volt)
F	focal point of a lens
	force
F	magnitude of a force
F	restoring force
F_{B}	buoyant force

Symbol	Definition
F_{c}	centripetal force
F_{i}	force input
$_{\text{net}}$	net force
F_{o}	force output
FM	frequency modulation
f	focal length
f	frequency
f_0	resonant frequency of a resistance, inductance, and capacitance (RLC) series circuit
f_0	threshold frequency for a particular material (photoelectric effect)

Symbol	Definition
f_1	fundamental
f_2	first overtone
f_3	second overtone
f_B	beat frequency
f_k	magnitude of kinetic friction
f_s	magnitude of static friction
G	gravitational constant
G	green quark color
G	antigreen (magenta) antiquark color

Symbol	Definition
g	acceleration due to gravity
g	gluons (carrier particles for strong nuclear force)
h	change in vertical position
h	height above some reference point
h	maximum height of a projectile
h	Planck's constant
hf	photon energy
h_i	height of the image
h_o	height of the object
I	electric current

Symbol	Definition
I	intensity
I	intensity of a transmitted wave
I	moment of inertia (also called rotational inertia)
I_0	intensity of a polarized wave before passing through a filter
I_{ave}	average intensity for a continuous sinusoidal electromagnetic wave
I_{rms}	average current
J	joule
J/Ψ	Joules/psi meson
K	kelvin
k	Boltzmann constant

Symbol	Definition
k	force constant of a spring
K_α	x rays created when an electron falls into an $n=1$ shell vacancy from the $n=3$ shell
K_β	x rays created when an electron falls into an $n=2$ shell vacancy from the $n=3$ shell
kcal	kilocalorie
KE	translational kinetic energy
$\mathrm{KE}+\mathrm{PE}$	mechanical energy
KE_e	kinetic energy of an ejected electron
$\mathrm{KE}_{\mathrm{rel}}$	relativistic kinetic energy
$\mathrm{KE}_{\mathrm{rot}}$	rotational kinetic energy
KE	thermal energy

Symbol	Definition
kg	kilogram (a fundamental SI unit of mass)
L	angular momentum
L	liter
L	magnitude of angular momentum
L	self-inductance
ℓ	angular momentum quantum number
L_α	x rays created when an electron falls into an $n = 2$ shell from the $n = 3$ shell
L_e	electron total family number
L_μ	muon family total number
L_τ	tau family total number

Symbol	Definition
L_{f}	heat of fusion
L_{f} and L_{v}	latent heat coefficients
$\mathrm{L}_{\mathrm{orb}}$	orbital angular momentum
L_{s}	heat of sublimation
L_{v}	heat of vaporization
L_z	*z* - component of the angular momentum
M	angular magnification
M	mutual inductance
m	indicates metastable state
m	magnification

Symbol	Definition
m	mass
m	mass of an object as measured by a person at rest relative to the object
m	meter (a fundamental SI unit of length)
m	order of interference
m	overall magnification (product of the individual magnifications)
$m\left({}^{A}\mathrm{X}\right)$	atomic mass of a nuclide
MA	mechanical advantage
m_{e}	magnification of the eyepiece
m_e	mass of the electron
m_ℓ	angular momentum projection quantum number

Symbol	**Definition**
m_n	mass of a neutron
m_{o}	magnification of the objective lens
mol	mole
m_p	mass of a proton
m_{s}	spin projection quantum number
N	magnitude of the normal force
N	newton
	normal force
N	number of neutrons
n	index of refraction

Symbol	Definition
n	number of free charges per unit volume
N_{A}	Avogadro's number
N_{r}	Reynolds number
$\mathrm{N} \cdot \mathrm{m}$	newton-meter (work-energy unit)
$\mathrm{N} \cdot \mathrm{m}$	newtons times meters (SI unit of torque)
OE	other energy
P	power
P	power of a lens
P	pressure
	momentum

Symbol	Definition
p	momentum magnitude
p	relativistic momentum
$_{\mathrm{tot}}$	total momentum
$_{\mathrm{tot}}$	total momentum some time later
P_{abs}	absolute pressure
P_{atm}	atmospheric pressure
P_{atm}	standard atmospheric pressure
PE	potential energy
$\mathrm{PE}_{\mathrm{el}}$	elastic potential energy
$\mathrm{PE}_{\mathrm{elec}}$	electric potential energy

Symbol	Definition
$\mathrm{PE_s}$	potential energy of a spring
P_{g}	gauge pressure
P_{in}	power consumption or input
P_{out}	useful power output going into useful work or a desired, form of energy
Q	latent heat
Q	net heat transferred into a system
Q	flow rate—volume per unit time flowing past a point
$+Q$	positive charge
$-Q$	negative charge

Symbol	Definition
q	electron charge
q_p	charge of a proton
q	test charge
QF	quality factor
R	activity, the rate of decay
R	radius of curvature of a spherical mirror
R	red quark color
R	antired (cyan) quark color
R	resistance
R	resultant or total displacement

Symbol	Definition
R	Rydberg constant
R	universal gas constant
r	distance from pivot point to the point where a force is applied
r	internal resistance
$r_{\perp}$	perpendicular lever arm
r	radius of a nucleus
r	radius of curvature
r	resistivity
r or rad	radiation dose unit
rem	roentgen equivalent man

Symbol	Definition
rad	radian
RBE	relative biological effectiveness
RC	resistor and capacitor circuit
rms	root mean square
r_n	radius of the *n*th H-atom orbit
R_{p}	total resistance of a parallel connection
R_{s}	total resistance of a series connection
R_{s}	Schwarzschild radius
S	entropy
	intrinsic spin (intrinsic angular momentum)

Symbol	Definition
S	magnitude of the intrinsic (internal) spin angular momentum
S	shear modulus
S	strangeness quantum number
s	quark flavor strange
s	second (fundamental SI unit of time)
s	spin quantum number
	total displacement
$\sec\theta$	secant
$\sin\theta$	sine
s_z	*z*-component of spin angular momentum

Symbol	Definition
T	period—time to complete one oscillation
T	temperature
T_c	critical temperature—temperature below which a material becomes a superconductor
T	tension
T	tesla (magnetic field strength *B*)
t	quark flavor top or truth
t	time
$t_{1/2}$	half-life—the time in which half of the original nuclei decay
$\tan\theta$	tangent
U	internal energy

Symbol	Definition
u	quark flavor up
u	unified atomic mass unit
	velocity of an object relative to an observer
	velocity relative to another observer
V	electric potential
V	terminal voltage
V	volt (unit)
V	volume
	relative velocity between two observers
v	speed of light in a material

Symbol	Definition
	velocity
	average fluid velocity
$V_{\mathrm{B}} - V_{\mathrm{A}}$	change in potential
$_{\mathrm{d}}$	drift velocity
V_{p}	transformer input voltage
V_{rms}	rms voltage
V_{s}	transformer output voltage
$_{\mathrm{tot}}$	total velocity
v_{w}	propagation speed of sound or other wave
$_{\mathrm{w}}$	wave velocity

Symbol	Definition
W	work
W	net work done by a system
W	watt
w	weight
w_{fl}	weight of the fluid displaced by an object
W_{c}	total work done by all conservative forces
W_{nc}	total work done by all nonconservative forces
W_{out}	useful work output
X	amplitude
X	symbol for an element

Symbol	Definition
${}^{Z}_{A}X_{N}$	notation for a particular nuclide
x	deformation or displacement from equilibrium
x	displacement of a spring from its undeformed position
x	horizontal axis
X_{C}	capacitive reactance
X_{L}	inductive reactance
x_{rms}	root mean square diffusion distance
y	vertical axis
Y	elastic modulus or Young's modulus
Z	atomic number (number of protons in a nucleus)

Symbol	Definition
Z	impedance